UNIVERSITY OF SOUTHAMPTON

Addressing Challenges in Gamified Paid Microtask Crowdsourcing Using Furtherance Incentives

by

Oluwaseyi Feyisetan

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Physical Sciences and Engineering
Electronics and Computer Science

September 15, 2016

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING
ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Oluwaseyi Feyisetan

Crowdsourcing has the potential to revolutionise the way organisations carry out tasks that need to scale out quickly – and indeed this revolution has begun. However, crowdsourcing today, and especially paid microtasks, face several technical and socio-economic challenges that can hamper the realisation of this vision. This work addresses four of such challenges: workflow design; real-time crowd work; motivation and rewards; and synchronous collaboration. The thesis describes the use of a bespoke gamified crowdsourcing platform *Wordsmith*, and studies the use of *furtherance incentives* to tackle issues at the heart of microtasks that feature monetary payments as the primary source of incentivisation. Furtherance incentives represent a timely and appropriate reward to improve task continuance presented when a worker is about to quit a task. As such, the keys to effectively deploying furtherance incentives lie in: the timely ability to detect waning worker interest in a task, and, knowledge of the appropriate incentive to offer the particular worker at that stage of the task.

In understanding how to improve crowdsourcing workflow designs, the thesis presents an approach that leverages on insights into task features and worker interaction preferences. The findings illustrate how workers interact with tasks in the presence of choice – thus offering us an idea into the types of furtherance incentive to offer workers. In the study on real-time crowd work, microtask contests are introduced as a medium to engage workers to complete tasks featuring tight time constraints. The results give us a rich model that we use to predict when workers are likely to exit a task at different stages. The research into motivation and rewards combines the two components of furtherance incentives by using gamification elements as an additional source of incentives. This leads to more tasks carried out and at a higher quality when compare with baseline paid microtasks. Finally our study on synchronous collaboration offers an additional case study on the effectiveness of furtherance incentives. Here we use sociality-based features of social pressure and social flow between interacting workers as furtherance incentives resulting in improved qualitative and quantitative results.

Declaration of Authorship

I, Oluwaseyi Feyisetan , declare that the thesis entitled and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as:
  - Feyisetan, Oluwaseyi, Elena Simperl, Ramine Tinati, Markus Luczak-Roesch, and Nigel Shadbolt. 'Quick-and-clean extraction of linked data entities from microblogs.' In Proceedings of the 10th International Conference on Semantic Systems, pp. 5-12. ACM, 2014.
  - Feyisetan, Oluwaseyi, Elena Simperl, Max Van Kleek, and Nigel Shadbolt. 'Improving paid microtasks through gamification and adaptive furtherance incentives.' In Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2015.

- Feyisetan, Oluwaseyi, Markus Luczak-Roesch, Elena Simperl, Ramine Tinati, and Nigel Shadbolt. 'Towards hybrid NER: a study of content and crowdsourcing related performance factors.' In The Semantic Web. Latest Advances and New Domains, pp. 525-540. Springer International Publishing, 2015.
- Feyisetan, Oluwaseyi, and Elena Simperl. 'Please Stay vs Let's Play: Social Pressure Incentives in Paid Collaborative Crowdsourcing' In The Proceedings of the 16th International Conference on Web Engineering, 2016.

Signed:...

Date:..

Acknowledgements

I would like to thank my supervisors Prof. Elena Simperl and Prof. Nigel Shadbolt, and all the members of the SOCIAM team in Southampton, Oxford and Edinburgh for their support towards the completion of this PhD thesis.

Contents

List of Figures

List of Tables

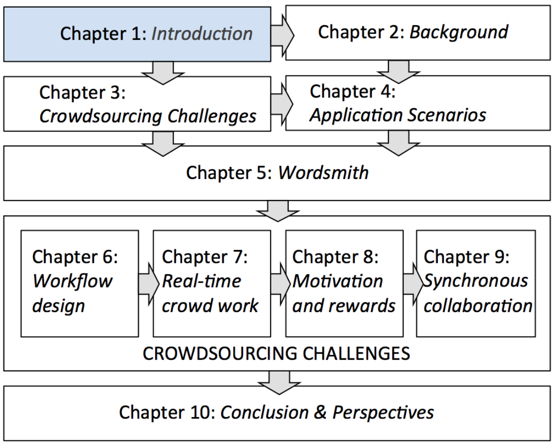
List of Abbreviations

AI	Artificial intelligence
API	Application Programming Interface
AV	Annotation Validation
CET	Cognitive Evaluation Theory
CEGE	Core Elements of the Gaming Experience
CRF	Conditional Random Fields
GATE	General Architecture for Text Engineering
GIT	General Interest Theory
GWAP	Games With a Purpose
HIT	Human Intelligence Tasks
IE	Information Extraction
IR	Information Retrieval
JSON	JavaScript Object Notation
MDA	Mechanics Dynamics Aesthetics
MICE	Money Ideology Coercion Excitement
MTurk	Mechanical Turk
NLP	Natural Language Processing
NER	Named Entity Recognition
NERD	Named Entity Recognition and Disambiguation
POS	Parts of Speech
SAPS	Status Access Power Stuff
SDT	Self Determination Theory
URI	Unique Resource Indicator
URL	Unique Resource Locator
WWW	World Wide Web

Dominus illuminatio mea

Chapter 1

Introduction



Chapter 1: *Introduction*
Chapter 2: *Background*
Chapter 3: *Crowdsourcing Challenges*
Chapter 4: *Application Scenarios*
Chapter 5: *Wordsmith*
Chapter 6: *Workflow design*
Chapter 7: *Real-time crowd work*
Chapter 8: *Motivation and rewards*
Chapter 9: *Synchronous collaboration*
CROWDSOURCING CHALLENGES
Chapter 10: *Conclusion & Perspectives*

This chapter presents an overview of our work. We begin with a high level introduction to paid microtasks. We then discuss our research questions, highlighting the four main challenges this thesis sets out to address. Subsequently, we list a summary of our major and additional contributions in line with the aforementioned challenges. The chapter is concluded with the organisation structure of the rest of the thesis.

1.1 Overview

Paid microtask crowdsourcing represent a new frontier in the way organisations carry out business - a faster, scalable and cheaper alternative to outsourcing. A single task is broken down into microtasks which can be solved rapidly in parallel by members of the 'crowd' who are recruited via an open call (Howe, 2006). Requesters and crowd workers meet in an online marketplace: the requesters post a task with instructions on how it is to be completed; and workers solve the tasks independently. The individual results are aggregated, the workers get paid and the requester gets their result. In most cases, only the requester has an overview of the purpose and scope of the entire task. Workers only see a micro-snippet e.g., a paragraph to annotate, a few seconds of transcription, or, a small block of text for translation. As such, there is a tendency to represent workers as human processors, one of many cogs in a wheel, or a member of a homogeneous set of low or static skilled personnel. The requesters could therefore see members of the crowd as replaceable pieces of machinery, leaving the workers with little incentive to carry out the task beyond the financial payout at the end of a few seconds of work. This does not result in the most engaging and rewarding work outlook - leading to widespread spamming and cheating on the side of the workers, and delayed or reneged payment on the part of the requesters. Poor incentivisation leads not only to bad task results, but also undermines the potential to carry out more complex workflows and cognitively demanding requests on crowdsourcing platforms.

This paints a bleak picture, which fortunately is not the case, as yet (Felstiner, 2011; Kittur et al., 2013; Martin et al., 2014). Economic forces have gradually pulled the average payout on crowdsourcing tasks steadily upwards over the last five years (Difallah et al., 2015), and researchers have been vocal about fair and ethical crowdsourcing (Irani and Silberman, 2013). This leads to higher financial incentives for workers to partake in crowd tasks. However, increased monetary payments is just a piece of the incentives puzzle and many challenges still remain. Microtask crowdsourcing has evolved beyond simple annotation tasks to complex workflows and creative tasks (Kittur et al., 2011; Kittur, 2010; Yu and Nickerson, 2011). In addition, crowd work sometimes yields better output than traditional individual work as it leverages on the wisdom of the crowd, forcing down prices as many workers vie for available tasks. This further increases its potential to displace workers in traditional employment as organisations seek to cut costs and optimise their output – thus leading to socio-economic questions and concerns akin to the more dystopian view of job losses from AI. Given this potential, it becomes necessary to design crowdsourcing platforms that are not only profitable to the requesters, but also provide a rewarding experience to the crowd workers.

1.2 Research Questions

Kittur et al. (2013) presented a research agenda on the future of crowd work that covered twelve points. In this thesis, we centre on four of the research foci in the context of paid microtasks that feature monetary payments, layered with other sorts of incentives.

RQ1. Workflow design - Can we understand what tasks (in hybrid human-machine workflows) are amenable to crowd work and thus route them accordingly?

RQ2. Real-time crowd work - Can we design systems that support timely worker recruitment and task execution on real-time work streams?

RQ3. Motivation and rewards - Can we leverage on gamification to build systems that are more rewarding and engaging than traditional crowdsourcing systems?

RQ4. Synchronous collaboration - Can we harness the power of collaboration to solve complex tasks?

These research questions are linked by an underlying commonality, which is our research methodology. We intend to address incentives as the root notion in tackling these crowdsourcing challenges. Our research hinges on the hypothesis that adopting a system of incentives – which we termed '*furtherance incentives*' – is vital in answering each of the individual questions. As is with crowdsourcing systems, we are also interested in keeping the financial costs bearable for the requesters and the task quality from the workers at an acceptable level. Equally important with these challenges, we seek to design systems that are appealing, engaging and rewarding for crowd workers. As the debate on ethical crowdsourcing continues (Irani and Silberman, 2013), we deem it necessary to tow the lines of fairness in our financial compensation strategies and place worker choice at the centrality of our experiments. For crowdsourcing to become a fully integrated and cost efficient source of solving complex tasks in organisations, **RQ1**, **RQ2** and **RQ4** serve as vital pieces of an already challenging puzzle. However, for microtask crowdsourcing to be judged as a 'morally' acceptable form of economic transaction between workers and a requester in the online world (given the tangentially related context of worker classification lawsuits in the offline world [1]), then addressing **RQ3** attracts commensurate consequence.

1.3 Summary of Contributions

The goal of this body of work is therefore to design methods and algorithms that address challenges in gamified paid microtask crowdsourcing. We implement an overarching crowdsourcing platform - *Wordsmith* - as a container for modules that tackle each of these challenges. Each module represents one of the research questions highlighted above, which in turn represents individual chapters in the thesis.

Below we detail our contributions and the associated publications.

RQ1. Workflow design

We demonstrate that crowd workers prefer and perform better on certain tasks when the element of choice is introduced. This allows us to design workflows that plays to the strength of the crowd, and thus either creates a refining/verification step in the crowdsourcing process or assigns difficult cases to a mediating team of experts. Our experiments focused on named entity recognition in microposts. Our results shed light on specific content features within the tasks, and behavioural features exhibited by the crowd that can help us decide what assignments would elicit the best quality response from workers.

[1] http://uberlawsuit.com/

Publication(s)

- Feyisetan, Oluwaseyi, Markus Luczak-Roesch, Elena Simperl, Ramine Tinati, and Nigel Shadbolt. 'Towards hybrid NER: a study of content and crowdsourcing related performance factors.' In The Semantic Web. Latest Advances and New Domains, pp. 525-540. Springer International Publishing, 2015.

- Feyisetan, Oluwaseyi, Markus Luczak-Roesch, Elena Simperl, Ramine Tinati, and Nigel Shadbolt. 'An extended study of content and crowdsourcing related performance factors in Named Entity Annotation' (*in submission*)

RQ2. Real-time crowd work

We explore the use of cardinal-ordinal contests as a platform for timely completion of real-time crowdsourcing tasks. Our approach allows us to carry out timely worker recruitment and collect judgements under tight time constraints with minimal loss of quality. We study the exit patterns of crowd workers to understand what point they drop out of contests. We use this to create an algorithmic predictive model to identify when workers are about to leave the task. The experiments here also focused on named entity recognition. We also demonstrate how contests can serve as an incentive beyond baseline financial payments.

Publication(s)

- Feyisetan, Oluwaseyi, and Elena Simperl. 'Performance and Exit Behaviour in Real-Time Crowdsourcing Contests' (*in submission*)

RQ3. Motivation and rewards

We illustrate the positive impact of gamification in improving the volume and quality of work undertaken in paid microtask settings. We also show that gamification leads to increased participation and engagement in tasks that had money as their primary incentive. We study various gamification elements to understand their impacts and we introduce the concept of furtherance incentives as a mechanism to prevent workers dropping off a task. We also create a predictive model to allocate the optimum furtherance incentive to prevent drop-off. The experiments here used simple image labelling tasks popular in the crowdsourcing literature.

Publication(s)

- Feyisetan, Oluwaseyi, Elena Simperl, Max Van Kleek, and Nigel Shadbolt. 'Improving paid microtasks through gamification and adaptive furtherance incentives.' In Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2015.

- Feyisetan, Oluwaseyi, and Elena Simperl. 'Incentives engineering in online crowdsourcing' (*in submission*)

RQ4. Synchronous collaboration

We demonstrate that workers performing paid microtasks would be willing to engage in collaborative workflows. We also illustrate how socially motivated incentives can lead to improved quality and volume of task output. Our results show that social pressure between workers and the desire to re-experience social flow, can serve as furtherance incentives to prevent workers dropping off the task. We provide evidence for empathic collaboration between workers; with workers willing to carry out extra tasks for free to ensure a fellow worker gets paid. The experiments here also used image-labelling tasks.

Publication(s)

- Feyisetan, Oluwaseyi, and Elena Simperl. 'Please Stay vs Let's Play: Social Pressure Incentives in Paid Collaborative Crowdsourcing' In The Proceedings of the 16th International Conference on Web Engineering, 2016.

- Feyisetan, Oluwaseyi, and Elena Simperl. 'Social Pressure Incentives in Paid Collaborative Crowdsourcing' (*in submission*)

1.4 Additional Contributions

In addition to the core contributions of this thesis, we also published work on automatic named entity recognition from large microposts datasets. This informed our study on **RQ1.** where we needed to design hybrid workflows that could support human and machine based annotations. The dataset created a repository from which we sampled out the corpus for our primary contributions.

Publication(s)

- Feyisetan, Oluwaseyi, Elena Simperl, Ramine Tinati, Markus Luczak-Roesch, and Nigel Shadbolt. 'Quick-and-clean extraction of linked data entities from microblogs.' In Proceedings of the 10th International Conference on Semantic Systems, pp. 5-12. ACM, 2014.

1.5 Outline

The remaining chapters of this thesis are organised as follows:

Chapter 2 – Background: in this chapter we give an overview on the concepts that run throughout the discourse of the thesis. We recall an historical rundown and synopsis of crowdsourcing, then highlight an overview of related socio-technical fields (such as human computation and collective intelligence) which influence our understanding of crowdsourcing. Afterwards, we shape our comprehension further by examining the dimensions of crowdsourcing before focusing specifically on paid microtask crowdsourcing. Subsequently, the chapter presents motivation and incentives, shedding light on money as a primary motivator before expounding on our additional incentive mechanisms. We introduce gamification as a foundational concept across our incentive studies and finally present background information on competitions and collaboration in crowdsourcing as these forms the basis for future specific chapters.

Chapter 3 – Crowdsourcing Challenges: this chapter gives an analysis of the challenges studied in the thesis. We present issues in workflow design; real-time crowd work; motivation and incentives engineering; and collaboration in paid microtask crowdsourcing. For each of the individual challenges, we describe the state of the art representing how the underlying issues are currently tackled. Given the potential scope of research questions that can be identified from each item, we outline the specific parts of the challenge that this work seeks to address.

Chapter 4 – Crowdsourcing Application Scenarios: in this chapter we explore two broad crowdsourcing application areas which form the basis of all our experiments in future chapters: text annotation and image labelling. The chapter also serves as a literature review on two front: (a) it presents related work in the line of our selected application scenarios and how crowdsourcing techniques have been applied; and (b) it discusses the state of the art in implementing customised platforms designed to address specific pain points encountered in crowdsourcing.

Chapter 5 – Wordsmith: in this chapter, we introduce *Wordsmith*, our bespoke crowdsourcing platform. We discuss its primary application areas vis-à-vis how it fits in with the challenges, scenarios and our experiment designs. We list the various modules that are used to address the different challenges and how Wordsmith's interface has been adapted to fit them. The chapter also describes the crowdsourcing process and how Wordsmith is designed to integrate through project definition, execution and quality control.

Chapter 6 – Workflow Design: in this chapter we describe the methods, experimental set-up, and data used to address the challenge of designing a useful workflow for crowdsourcing named entities. We discuss the potential of building better workflows for paid microtasks by leveraging on insights into task features and worker preferences. This chapter also introduces the concept of furtherance incentives, which is expanded in later chapters. It presents a conceptual approach on how our findings can be used to re-integrate worker preferences as an incentive mechanism.

Chapter 7 – Real-time Crowd Work: in this chapter we address a specific challenge in real-time crowd work by using crowdsourcing contests in combination with individual micro-payments to collect judgements effectively under tight time constraints. We present our crowdsourcing contest model followed by our approach at predicting worker drop-offs. Following from the previous chapter, we continue our discourse on furtherance incentives by expounding on the components that afford for the deployment of furtherance incentives, one of which is predicting potential task exits.

Chapter 8 – Motivation and Rewards: in this chapter, we build upon, and come full circle on the concept of furtherance incentives. This chapter examines the potential of adding gamification to microtask interfaces as a means of improving both worker engagement and effectiveness. It also defines a predictive model for estimating the most appropriate furtherance incentive for individual workers, based on their previous contributions. This allows us to build a personalised game experience, with gains seen on the volume and quality of work completed.

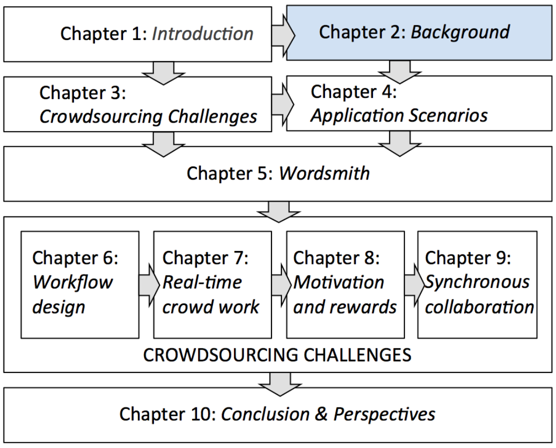
Chapter 9 – Synchronous Collaboration: in this chapter, we address the fourth challenge of synchronous collaboration. We also apply the concept of furtherance incentives in continuance with the insights gained from the previous chapters. In particular, results from our study on motivation and rewards indicated sociality based incentives were the most effective drivers of retention and engagement. We therefore expand our knowledge by experimenting with two sociality-driven furtherance incentives – social pressure and social flow in our study of synchronous collaboration in microtask crowdsourcing.

Chapter 10 – Conclusions and Perspectives: We close this thesis by presenting a summary of the work done and the contributions made. We also outline future work in addressing further research areas in paid microtask crowdsourcing and the potentials of applying furtherance incentives to tackle other issues in crowdsourcing.

Within the thesis, each chapter is preceded by a summary of the chapter highlighting what it is set to achieve. At the end of each chapter, we also revisit the contributions made and the lessons learnt.

Chapter 2

Background


Chapter 1: *Introduction*
Chapter 2: *Background*
Chapter 3: *Crowdsourcing Challenges*
Chapter 4: *Application Scenarios*
Chapter 5: *Wordsmith*
Chapter 6: *Workflow design*
Chapter 7: *Real-time crowd work*
Chapter 8: *Motivation and rewards*
Chapter 9: *Synchronous collaboration*
CROWDSOURCING CHALLENGES
Chapter 10: *Conclusion & Perspectives*

In this chapter we present background material that form the foundations of the thesis. We begin with an historical introduction to 'crowdsourcing' in general. Next, we give an overview of related fields that influence our understanding of crowdsourcing. Afterwards, we shape our comprehension further by examining the dimensions of crowdsourcing before focusing specifically on paid microtask crowdsourcing. Finally, we introduce the subject of motivation in paid microtasks, beginning with the base motivator of monetary payments, concluding the chapter with additional incentive mechanisms that make up constituent components of future chapters.

2.1 Introduction

In 1785 (de Caritat et al., 1785), the Marquis de Condorcet proposed a theory about the probability of a collective group of error prone decision makers correctly coming to the right choice on one of two decisions. This theorem, the *Condorcet's jury theorem*, represents one of the earliest postulations on how performance can increase given a large enough crowd i.e., as the number of people grow, the probability of choosing the right answer approaches 1. In 1906, Sir Francis Galton (Galton, 1907) carried out an experiment which demonstrated what he described as *Vox Populi* – the Wisdom of the Crowd. A group of 800 people, including butchers and farmers, were drawn into a competition to guess the weight of a live ox. The final average weight from 787 of them resulted in a value that was within 0.8% of the correct answer. Despite the contestants including those who were termed as 'highly experts at judging the weight of cattle', no single individual got the right answer, and the average deviation was between -3.7% and +2.4% of the actual answer. This ideology was rehashed by Surowiecki (2004) in a book with a title akin to Galton's. He identified instances where collective intelligence trumped individual smartness: for example, in the game show 'who wants to be a millionaire', the audience was right 91% of the time, while the individual expert was right only 65% of the time.

In creating a historical context for human computation, crowdsourcing, and leveraging the wisdom of the crowd, we discover that both the principle and practice date back to the 18th century. In 1783, King Louis XVI of France made an open call for a better way to produce alkali from sea salt, which was won by Nicolas Leblanc in 1791. In 1714, the Longitude Prize was offered for a practical method to determine a ship's longitude at sea (Moldovanu and Sela, 2001). These two inducement prizes took advantage of the power of the many to tackle difficult problems. The creation of the Oxford Dictionary in the 1800s also relied on thousands of volunteers from schools and universities, contributing quotations over four decades. However, these were groups of people working as individuals. Their output was devoid of a notion of the human and the mechanical working in tandem to achieve a common goal, neither was there an automated way to coordinate their contributions. In 1769, Baron Wolfgang von Kempelen built the first machine which could, purportedly beat humans at chess (Levitt, 2000). For over 80 years, 'The Mechanical Turk', as it was called beat several humans at chess – including Napoleon and Benjamin Franklin. The Turk finally turned out to be a hoax - a clever piece of machinery, housing an actual chess master. During its time, The Turk was operated by at least 6 different chess masters – sourcing intelligence from an expert crowd. The Turk or Automaton Chess Player could probably be termed an early ancestor of Deep Blue, the computer that went on to actually defeat a human grandmaster, however, the term 'computer' also once referred to a job description of actual humans carrying out computations. During World War II, there were a number of areas where extensive, continuous calculations were needed, and, a large pool of women with training in mathematics (Erickson et al., 2010). These women became the early 'computers', physically cranking out calculations in what was to be the precursor to the Electronic Numerical Integrator and Computer (ENIAC) – one of the first electronic computers.

By the time Jeff Howe published his article on 'The Rise of Crowdsourcing' in 2006 (Howe, 2006), the groundwork was already laid and ideas had started taking steam; computers were getting more powerful and the web was part of everyday business. The article chronicled crowdsourcing systems such as: Threadless, which crowdsourced shirt designs; iStockphoto, a disruptive marketplace for amateur photographers; and InnoCentive, an ideas hub which cut R&D budgets by leveraging on the wisdom of the crowd. However, what would turn out to be the leading crowdsourcing platform was still in its infancy: described as a work in progress without an official launch date – even though it had been re-inventing how businesses carried out their activities. In 2005, Amazon launched Mechanical Turk – named after von Kempelen's automaton. It was described as 'artificial artificial intelligence' – humans working behind the scenes to solve HITs (Human Intelligence Tasks) that could not presently be carried out by computers. Today, after over 10 years, Amazon's Mechanical Turk (or MTurk) has tens of thousands of 'turkers' carrying out HITs daily, over 1,000 new requesters joining per month and north of 130 million HITs created in the years from 2009 to 2014 (Difallah et al., 2015; Ipeirotis, 2010a). MTurk represents a class of crowdsourcing – paid microtask crowdsourcing – powered by large-scale access to the Internet and the ease of online payment transactions, serving as a source of cheap (and possibly high quality) data (Buhrmester et al., 2011). The speed and reach of the Internet makes it possible for a project like Wikipedia to be undertaken within a time-frame significantly shorter than the Oxford Dictionary project. What began organically over the centuries as people sought better ways to harness collective intelligence has now become a global phenomenon powered by access to the web.

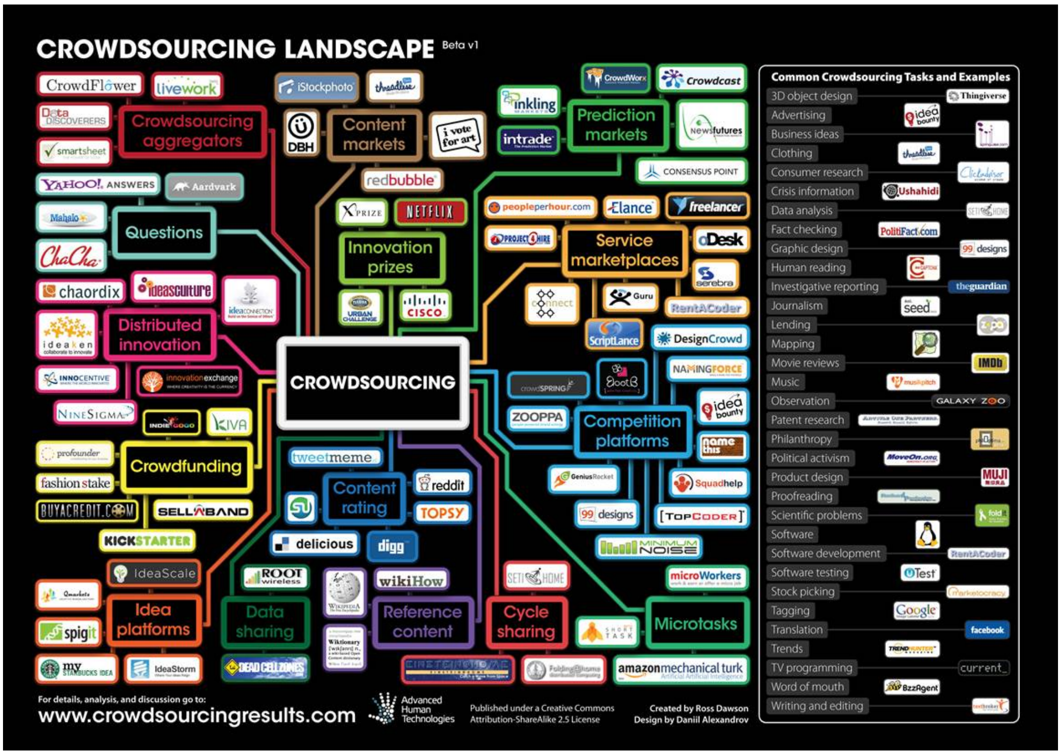


FIGURE 2.1: Crowdsourcing Landscape (crowdsourcingresults.com)

The web today serves as a symbol of information democratisation and the continued lowering of barriers to entry within fields that were once siloed in walled enterprises. Crowdsourcing via the web has found its way gradually over the past 10 years into the mainstream data pipeline of industries ranging from high technology and government agencies to the military and sales agencies. What once started as a phenomenon described in the physical world, now finds its quintessential expressions on the web. Crowdsourcing, though riddled with its challenges (Kittur et al., 2013) from the technical to the ethical is here to stay. Even as it morphs continually to the tune of academic and economic forces, crowdsourcing remains a force for good with ground-breaking discoveries being made daily directly and indirectly through the wisdom of the crowd (Cooper et al., 2010; Kim et al., 2014). Figure 2.1 shows a rich mix of crowdsourcing site types featuring ideas platforms, innovation prizes, crowdfunding, content and prediction markets, competition platforms and service marketplaces. The figure also shows popular crowdsourcing tasks and their corresponding example. Paid microtask crowdsourcing – identified by its poster child - Amazon's Mechanical Turk represents but a small subset of the landscape but has grown to be a very vital component in the crowdsourcing economy. Similarly, Crowdsourcing.org – a website that bills itself as an industry resource on crowdsourcing topics, aggregates crowdsourcing platforms into 5 categories: [1] cloud labour, crowd creativity, crowdfunding, distributed knowledge, open innovation and tools. The directory currently holds close to 3,000 sites (although a large portion i.e. over 800 are within the crowdfunding and distributed knowledge space). Microtask crowdsourcing, comprising about 135 sites, sits in the cloud labour category along side expert tasks and freelancing platforms.

2.2 Related Fields

According to Howe (2006) who is credited for coining the term, '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'. Related research include Human Computation – (Quinn and Bederson, 2011), Social Computing – (Parameswaran and Whinston, 2007), Collective Intelligence – (Weiss, 2005), Social Machines – (Smart et al., 2014) and Technology-Mediated Social-Participation Systems – (Kraut et al., 2010). While these systems are all similar, with significant overlapping exemplars, each one has managed to carve out a niche that makes it slightly different from the others. Figure 2.2 by Shadbolt et al. (2013) shows the relationship and overlaps between crowdsourcing and other related fields.

2.2.1 Human Computation

Von Ahn (2005) in his 2005 doctoral dissertation defined human computation as 'a paradigm for utilising human processing power to solve problems that computers cannot yet solve'. According to Quinn and Bederson (2011) with human computation, the target problems are computational (e.g., image and speech recognition) – which might or might

[1] http://www.crowdsourcing.org/directory

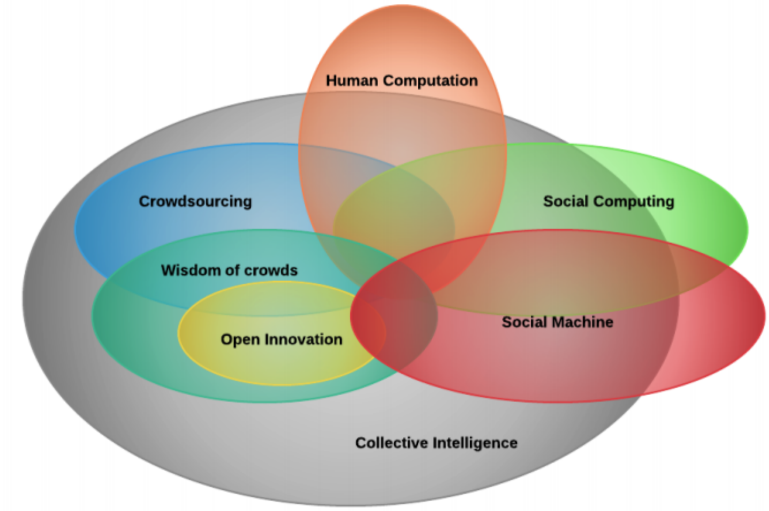


FIGURE 2.2: Crowdsourcing and related areas by Shadbolt et al. (2013)

not be the case with crowdsourcing. As such human computation generally replaces computers with humans while crowdsourcing replaces trained experts with members of the crowd. This of course is a loose generalisation and indeed, there exists overlaps between crowdsourcing and human computation in instances where crowd workers act as computation processors (e.g., in translation and text annotation tasks). Quinn and Bederson (2011) presented a survey on human computation as well as a summary of definitions from literature. Other fields related to crowdsourcing also find considerable overlaps with human computation – as can be seen from Figure 2.2.

A unique sub-class of human computation systems are called Games With a Purpose (GWAP) (von Ahn and Dabbish, 2008). Unlike traditional crowdsourcing (and specifically paid microtask crowdsourcing), participation in these systems is not motivated by financial gain, but by the pleasure and enjoyment derived from playing a game. The computational output of the players is usually derived as a side effect of gameplay. This concept has been integrated into crowdsourcing systems via gamification (Zichermann, 2011; Feyisetan et al., 2015b) i.e., the use of game mechanics in non-game context. Examples of GWAPs include Fold.it (a protein folding game) and the ESP Image Tagger (von Ahn and Dabbish, 2004).

The ESP Image Tagger by von Ahn and Dabbish (2004) (perhaps the most popular GWAP, and a model for crowdsourcing image labels) is a human computation styled game used to collect keywords which can suitably describe an image. The game was designed to harness the instrumentation of human workers to address a problem, which was difficult (at the time) to be handled by computers. The underlying principle was, by presenting a task as a game that is both fun and interactive, people would desire to play while generating useful output. In the ESP Image Tagger, randomly co-assigned paired players are presented with an image. The objective is to guess what keyword the other player is typing which depicts the image seen. At any point of agreement, the players are advanced onto the next image. Words which both players match on, and indeed which multiple game player pairs agree on can be presented with certain confidence to represent characteristic words for the given image.

2.2.2 Social Computing

Definitions of social computing tend to fall into two broad camps with each one leaning either more to the social or computing side. As noted by Robertson and Giunchiglia (2013), social computing has historically been used in a broad sense to describe socio-technical problem solving. Within one sense, such as presented by Ali-Hassan and Nevo (2009), social computing involves tools that facilitate social interactions such as blogs, wikis and social networks. This tends to compass technology-mediated communication processes, which are usually void of computation. In these systems, the natural social activities of the users, such as information exchange (e.g., via rich multimedia) take center stage (as against a primary computational requirement in crowdsourcing or human computation).

A more technical view of social computing such as by Robertson and Giunchiglia (2013), views social computing as consisting of 'programs that depend on algorithms that must run in human society in consonance with the computer systems'. This definition covers systems like online auctions and prediction markets (e.g., the Hollywood Stock Exchange).

2.2.3 Open Innovation

Open Innovation stands in stark contrast to the traditional notion of closed innovation where companies run their internal research and development teams within their corporate boundaries. Chesbrough (2006) describes open innovation as: 'the use of purposive inflows and outflows of knowledge to accelerate internal innovation, and expand the markets for external use of innovation, respectively'. This paradigm assumes that firms can and should use external ideas as well as internal ideas, and internal and external paths to market, as they look to advance their technology. Similar to crowdsourcing, open innovation sources the wisdom of the crowd via an open call. Indeed some authors such as Seltzer and Mahmoudi (2012) class open innovation systems as crowdsourcing platforms. However, for our discourse, open innovation relates to tasks solved as a whole unit at the macro level (e.g., a proposal to develop acetone-responsive materials) as against decomposable crowdsourcing tasks solved in parallel by multiple workers. In addition, solving open innovation challenges requires specialised knowledge or expertise in the requester's domain; as against traditional microtask crowdsourcing where the scale of workers required usually precludes the need for experts, and contribution from the crowd can be reduced to repetition within narrowly prescribed bounds (Seltzer and Mahmoudi, 2012).

Chesbrough (2006) presented an extensive look at Open Innovation with his eponymous book. Open innovation as a socio-technical system had its origins in business and management even though the primary examples of closed and open innovation presented by Chesbrough were technology companies. Today, systems such as InnoCentive serve as crowdsourcing platforms for Open Innovation, thus yielding a technology-mediated base. Open Innovation therefore presents numerous advantages (by leveraging the wisdom of external crowds) such as a trimmed down cost of research and development, additional source of free (and potentially viral) marketing of a brand and increased customer targeting. This however does not keep out challenges unique to human systems, in this case, the potential loss of a competitive advantage by wrongfully disclosing information. However, 'open systems', such as Tesla Motors practice of open sourcing patents (Musk, 2014) and the Open Compute Project by Facebook [2] has proved the benefits of accessible knowledge in an industry. Open Innovation, powered by humans on a technology platform, thus stands as a precursor to these nascent open movements.

2.2.4 Collective Intelligence

This term covers the most expansive area when compared with other ideas in related fields. Almost all instances of socio-technical systems described above can be termed collective intelligence systems in a broad sense. For example, Mataric (1993) described collective intelligence as a social construct emerging from individual intelligence, thus extending as far out as collective animal intelligence. Most modern systems termed as collective intelligence systems are web based e.g., Wikipedia. However, some collective intelligence projects such as open source developments, do not yield a web based service even though they are built collaboratively with online tools (e.g., github). Examples of these include the Linux and Apache Software Projects. Malone et al. (2010) presented a report on the Collective Intelligence Genome which introduces a taxonomy of these systems. Also within the collective intelligence space is the notion of the 'Wisdom of the Crowd' by Surowiecki (2004), which is applied in non web based scenarios such as collective decision making in organisations.

Several other systems can be classed under the umbrella of Collective Intelligence. These include systems that rely on implicit crowd knowledge or activity, and explicit machine computation. An example of this would be recommender systems (Resnick and Varian, 1997) such as those on eCommerce sites (e.g., Amazon and eBay). These leverage on our day to day experiences of making choices based on recommendations from other people either by word of mouth, movie reviews, book reviews etc. Recommender systems work by amplifying these natural social processes on web platforms by aggregating user input

[2] Open Compute Project - `http://www.opencompute.org/`

choices, and making suggestions to recipients. The suggestions could be presented based on relationship degrees of separation or historically similar choices.

Other implicit crowd platforms include social network email filters (Charles, 2010) and spam/ham classifiers which rely on either a user's social graph, or on training data provided by people. Click ranking, upvoting and several search result rankings also depend on most popular links clicked by users. Figure 2.2 by Shadbolt et al. (2013) shows the subsuming relationship and overlaps between collective int and other related fields.

2.2.5 Social Machines

In the book 'Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web' by Berners-Lee and Fischetti (1999), Berners-Lee and Fischetti state that,

> Real life is and must be full of all kinds of social constraint, the very processes from which society arises. Computers can help if we use them to create abstract social machines on the Web: processes in which the people do the creative work and the machine does the administration.

An immediate observation is the clear demarcation of task types – which brings to mind a hybrid system such as von Kempelen's Automaton Chess Player (Levitt, 2000) – humans and machines working in tandem to achieve a common goal. A definition which extends on Berners-Lee and Fischetti, addressing the limitations imposed on the constituent actions performed by the composing elements of a Social Machine has been offered by Smart and Shadbolt (2014). According to them: Social machines are web based socio-technical systems in which the human and technological elements play the role of participant machinery with respect to the mechanistic realisation of system level processes. Another definition by the Social Machines Lab at MIT [3] casts social machines as 'digitally mediated human networks – governed by AI and machine learning'.

Social machines therefore encompass a broad spectrum of systems from similar instances such as GalaxyZoo [4] and PlanetHunters [5] (both part of the Zooniverse Citizen Science Project) to reCAPTCHA [6] and LinkedIn [7]. Several social machines are also mentioned as examples in Social Computing, Crowdsourcing and Human Computation. Smart and Shadbolt (2014) include Wikipedia, social networks (Facebook, mySpace), Twitter and YouTube. Others [8] further note systems as diverse as Uber, Airbnb, Netflix and

[3] http://socialmachines.media.mit.edu/
[4] http://www.galaxyzoo.org
[5] http://www.planethunters.org
[6] http://www.google.com/recaptcha
[7] http://www.linkedin.com
[8] socialmachines.media.mit.edu

Snapchat, as examples of social machines. An attempt to create a bounding taxonomic framework for social machines was carried out by Smart et al. (2014).

2.3 Crowdsourcing Dimensions

To leverage on the wisdom of the crowd, either in human computation, collective intelligence or crowdsourcing systems, Surowiecki (2004) posits that deriving benefits from the crowd requires the following: a diverse group, with minimal cross decision influence, and a democratic process of maintaining quality. This serves as an initial pointer to dimensions of crowdsourcing i.e., what to think about and what might constitute a crowdsourcing campaign. Howe (2008) presented 10 of such rules to consider. Each rule reveals the potential for different dimensions to crowdsourcing. For example, the first 3 state: pick the right model, the right crowd and the right incentive – hinting at the existence of different models, crowds and incentive schemes.

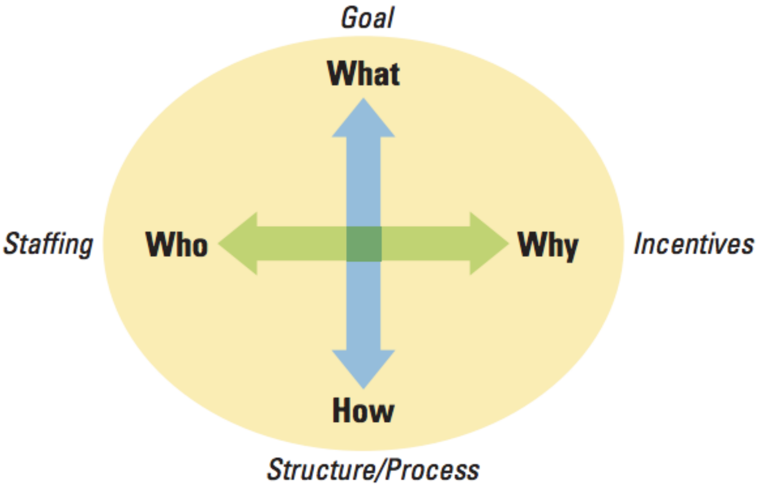


Figure 2.3: Crowdsourcing dimensions adapted from the Collective Intelligence Genome by Malone et al. (2010)

1. Pick the right model
2. Pick the right crowd
3. Offer the right incentives
4. Keep employing people
5. Find benevolent dictators
6. Keep things simple
7. Be prepared for fluff
8. Look for diamonds in the rough
9. The community is always right
10. Give the crowd something

Taxonomy (the practice of classification along dimensions) and taxonomies have surfaced in crowdsourcing and similar domains including: a categorisation of crowdsourcing systems on the web by Doan et al. (2011), a survey and taxonomy of human computation systems by Quinn and Bederson (2011), the Collective Intelligence Genome report by Malone et al. (2010), taxonomic dimensions in social computing by Ali-Hassan and Nevo (2009) and a taxonomic framework for social machines (Smart et al., 2014). The Collective Intelligence Genome report by Malone et al. (2010) and as shown in Figure 2.3 presents a simple 4 step approach which we can re-purpose to reason about crowdsourcing dimensions:

1. What (goals)
2. Who (staffing)
3. Why (motivation)
4. How (structure/process)

Dimension	Values / Characteristics	Example
What	Create	creating a new a e.g., artefact, transcription, summarization
	Decide	evaluating existing material e.g., reviewing, verification
Who	Crowd	parallelizable tasks to be solved independently
	Experts	tasks that create ambiguity with no crowd consensus
Why	Money (Pay)	microtask crowdsourcing e.g., MTurk, CrowdFlower
	Love (Enjoyment)	gamified systems e.g., The ESP Game, Phetch
	Glory (Reputation)	crowdsourcing contest platforms e.g., TopCoder
How - Create	Collection	parallelizable tasks to be solved independently
	Contest	tasks requiring a few correct answers
	Collaboration	tasks requiring continuous consensus
How - Decide	Group decision	everyone in the group needs to abide by the same decision
	Voting	important for the crowd to be committed to the decision
	Averaging	crowd has no systematic bias about estimating the number
	Consensus	group is small enough or has similar views

Table 2.1: Crowdsourcing dimensions adapted from Malone et al. (2010)

2.3.1 What

The '***what***' refers to what is outsourced to the crowd and what is the goal to be achieved. This highlights tasks suitable for crowdsourcing i.e., processes based on human skill which cannot be satisfactorily undertaken by machines. This is related to the 'human skill' dimension in the human computation taxonomy by Quinn and Bederson (2011) which lists visual recognition, language understanding and basic human communication as some aspects where humans traditionally excel beyond machines – thus representing suitable task types for crowdsourcing. Malone et al. (2010) divides what is being done into two: *creating* and *deciding*. This is identical to two of the dimensions presented by Doan et al. (2011) as *building artefacts* and *evaluating things*. When creating or building, workers generate something new: such as textual knowledge (e.g., a Wikipedia entry), structured knowledge (e.g., Wikipedia infoboxes) or a piece of software (e.g., Apache projects). On the other hand, workers decide, evaluate or refine tasks by voting, tagging or reviewing existing material.

Gadiraju et al. (2014) presented six top-level classes of tasks which are carried out on crowdsourcing platforms. This gives insight into the '*what*' of crowdsourcing:

1. **Information finding** (IF): refers to tasks that require workers to source information on the web for a specific question. For example, 'find the names of all the c-level executives in a list of companies'.

2. **Verification and validation** (VV): refers to tasks that require content verification, content validation, spam detection and data matching. For example, validating web domains or checking Twitter for spam accounts.

3. **Interpretation and analysis** (IA): leverages on the subjective wisdom of the crowd to carry out classification, categorization, ranking, sentiment analysis, content moderation and quality assessment – which are areas human are good at.

4. **Content creation** (CC): involves tasks that require workers to generate new content such as transcription an audio file, translation a document, text summarization, data enhancement and data annotation.

5. **Surveys** (SU): refers to tasks that require workers to give their opinion on a product, or give feedback on a service. It also includes collection of customer satisfaction data or demographic information.

6. **Content access** (CA): requires workers to access some content. For example, workers might be required to watch an online video. This could be to test a service or to promote its presence.

A longitudinal analysis carried out by Difallah et al. (2015) on the six top-level task types discussed in Gadiraju et al. (2014) (and illustrated in Figure 2.4) showed that

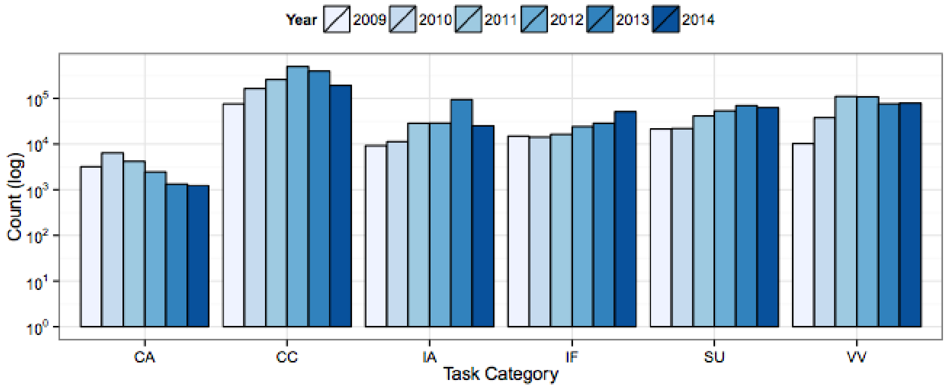


Figure 2.4: Popularity of HIT types over time by Difallah et al. (2015)

content creation tasks (such as data annotation and transcription) are the most popular while content access tasks are the least requested.

The social machines framework by Smart et al. (2014) give further insight into decomposing the 'what' of crowdsourcing. They list a set of dimensions and characteristics for goals, tasks and processes in social machines. Of importance here are the characteristics that relate to goals and tasks in crowdsourcing i.e., the goal variability and visibility. The goal variability refers to the possibility of the task goal to change. Most crowdsourcing goals are fixed per task for each user, however, in complex workflows such as the *find-fix-verify* model by Bernstein et al. (2010), workers might perform a single task with multiple objectives. For example, a worker might annotate an image, then verify an annotation on another. The goal visibility is usually hidden from the workers as a unit task e.g., a paragraph to be translated, does not reveal the entire scope and goal of the task.

2.3.2 Who

The '***who***' represents the type of crowd. This corresponds to rule 2/10 presented by Howe (2008) above. Although the crowd is sourced via an open call to an anonymous set of people, several platforms (e.g. CrowdFlower and Mechanical Turk) offer some degree of profiling e.g., based on language skills or geography. The profiling offers limited insight compared to what might be obtainable in traditional work settings, however, being able to target a large set of speakers of a particular esoteric language for example, is sufficient in specialised requests such as translation or annotation. Some crowdsourcing tasks might also require a pool of people with a certain expertise – an approach described by De Boer et al. (2012) as nichesourcing. Other crowdsourcing requests ascertain the suitability of the crowd by using qualifying questions to determine skill level. Finally, crowdsourcing systems also track the ratings and reputation of workers, allowing requesters to target their tasks to workers who have consistently produced high quality output.

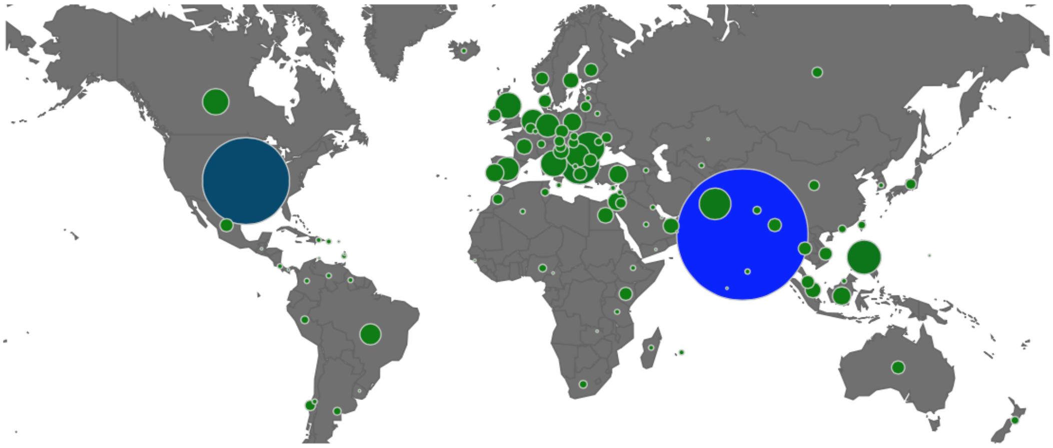
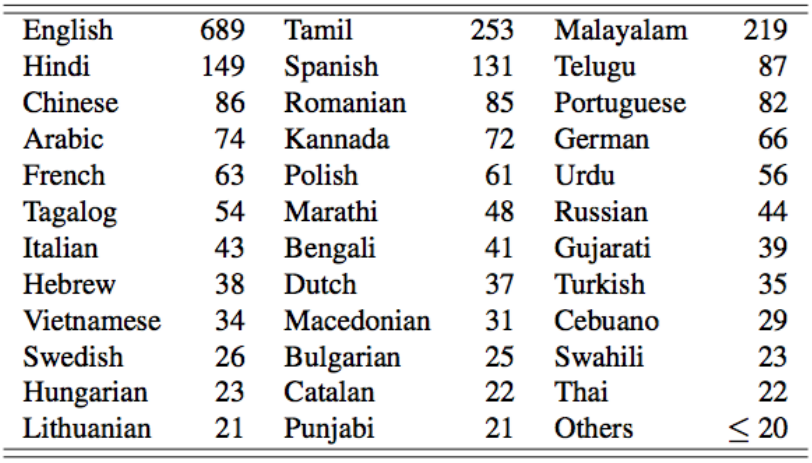


Figure 2.5: The number of crowd workers per country by Pavlick et al. (2014)

Figure 2.5 shows the number of crowd workers per country from a study by Pavlick et al. (2014). The results geo-locate 4,983 workers with the circles representing the number of workers in each country. The countries with the two highest worker counts are India and the United States. Other demographic studies by Ross et al. (2010) and Ipeirotis (2010b) shed more light on the characteristics of the two dominant countries (India and the United States) by reporting findings on the gender distribution, age and occupation of workers as well as their derived income. Difallah et al. (2015) also reported analysis which shows that certain tasks are restricted on a geographic basis – for example, surveys are assigned mostly to workers from the United States.



English	689	Tamil	253	Malayalam	219
Hindi	149	Spanish	131	Telugu	87
Chinese	86	Romanian	85	Portuguese	82
Arabic	74	Kannada	72	German	66
French	63	Polish	61	Urdu	56
Tagalog	54	Marathi	48	Russian	44
Italian	43	Bengali	41	Gujarati	39
Hebrew	38	Dutch	37	Turkish	35
Vietnamese	34	Macedonian	31	Cebuano	29
Swedish	26	Bulgarian	25	Swahili	23
Hungarian	23	Catalan	22	Thai	22
Lithuanian	21	Punjabi	21	Others	≤ 20

Figure 2.6: Self-reported native language of 3,216 bilingual Turkers by Pavlick et al. (2014)

Pavlick et al. (2014) further presented demographic results (illustrated in Figure 2.6) on the self-reported native language of 3,216 bilingual workers on Mechanical Turk. Their results indicate at least 35 languages spoken by over 20 workers each.

2.3.3 Why

The '***why***' attempts to answer the question: why would anyone be willing to take part as a worker in a crowdsourcing exercise. This corresponds to rule 3/10 (incentivising the crowd) and rule 10/10 (keeping the crowd motivated) presented by Howe (2008) above. Addressing the motivations of the crowd serves as a research topic in its own right and we would attempt to give a brief introduction here, to be expanded in further sections beginning from section 2.4.1 below.

Malone et al. (2010) and Quinn and Bederson (2011) both mention money/pay and enjoyment/love as two motivating factors for working on crowdsourcing tasks. However, in their paper aptly titled 'more than fun and money', Kaufmann et al. (2011) argue that motivations of crowd workers extend beyond monetary payment and the desire to be entertained. Modern motivation theories have evolved from the initial models by Deci and Ryan (1985b) to encompass richer models built on *intrinsic* and *extrinsic* motivation.

Even though it might appear that 'financial payouts' aptly represent extrinsic motivations, and the desire to have 'fun' fit intrinsic motivation, Kaufmann et al. (2011) and others such as Archak (2010) and Rogstadius et al. (2011), present the motivation of the crowd in terms of more complex parameters.

2.3.4 How

The '***how***', closely related to the '*what*', refers to the processes of carrying out the crowdsourcing task. This depends on whether the task refers to *creating* or *evaluating* something. The Collective Intelligence genome by Malone et al. (2010) highlights 3 methods of 'creating': (a) collection, (b) collaboration and (c) contests. These were the 3 approaches we adopted in our experiments detailed from Chapter 6 to 9.

1. **Collection**: In a *collection* approach to crowdsourcing, a task is broken down into small parallelizable pieces of work which can be carried out independently by multiple workers. For example, to create a volume of 1,000 articles, or to annotate 1,000 images (as in our experiments), an individual worker might be assigned to undertake just between 1-10 pieces of work. Chapter 6 and 8 employ a collection approach to crowdsourcing.

2. **Collaboration**: With *collaboration*, multiple workers are assigned to carry out a unit task i.e., rather than a single worker completing a piece of work, several people either pass the task around for incremental refinement, or work towards consensus. For example, Kittur (2010) describes a translation task collaboratively carried out by multiple workers. We describe a collaborative approach in detail in Chapter 9 while discussing the challenge of synchronous collaboration in crowdsourcing.

3. **Contest**: Another approach to the 'how' of crowdsourcing is to use *contests*. Contests are used either when only a few good results are required, e.g., the Netflix prize to create a better recommendation algorithm (Bennett and Lanning, 2007); or when an aggregated task needs to be solved as quickly as possible e.g., crowdsourcing for disaster relief. We describe a contest approach to addressing the challenge of real-time crowdsourcing in Chapter 7.

These three categories describe the process in crowdsourcing tasks that 'create new artefacts. However, some other tasks simple decide, refine or evaluate existing results. Similarly, some 'creation tasks might require an intermediary refinement step to evaluate or aggregate results before the final output is produced. Methods used to achieve this include voting, averaging and by consensus.

2.3.5 Quality Control

Besides the four dimensions discussed, i.e. goals (what), staffing (who), motivation (why) and structure/process(how), Smart et al. (2014), Allahbakhsh et al. (2013) and Quinn and Bederson (2011) present an additional dimension which is especially pertinent to crowdsourcing systems. This is the quality control or quality assurance mechanism. Cheating and spamming were reported by Difallah et al. (2012) as one of the main issues with crowdsourcing. This is especially the case in paid microtask crowdsourcing (properly introduced in Section 2.4 below) where the primary incentive mechanism is financial payments. They reported several adversarial techniques employed by individual malicious workers to bypass the task requirements and receive the monetary reward. This includes submitting random, automated or semi-automated answers. This submissions are usually either artificially generated, or duplicated from an existing task. A group of workers could also collude maliciously in tasks that require consensus thus leading to false agreements on answers or answer sharing.

Apart from dishonest submissions, quality control is also essential to accommodate for answers submitted as a result of workers not fully comprehending the task instructions. Some approaches to addressing quality control issues in paid microtask crowdsourcing adapted from Quinn and Bederson (2011) and Allahbakhsh et al. (2013) include:

- **Redundancy**: this is one of the most common approaches which comes by design with crowdsourcing platforms such as Mechanical Turk and CrowdFlower. By employing multiple workers, and based on the research by Difallah et al. (2012) which shows that malicious workers are in the minority, requesters can use a voting system to identify possibly correct answers while weeding out workers who consistently diverge from the wisdom of the crowd.

- **Output agreement**: In tasks that rely on output agreement, workers are paired (at the minimum), and their responses are only accepted as valid if they attain joint consensus. This is the approach popularised by von Ahn and Dabbish (2004) in the ESP game. Output agreement requires synchronicity between workers to advance between tasks. Carvalho et al. (2014) also showed that output agreement could induce honest behaviour when workers believe their submission is the correct one which other workers must agree with.

- **Multi-level review**: With multilevel review, a unit task goes through a sequential set of incremental refinement or verification processes before the final output is produced. This approach is employed in workflows such as '*find-fix-verify* by Bernstein et al. (2010) where the output of one worker becomes the input for the next. Multilevel review can occur synchronously, in a collaborative fashion similar to output agreement such as described by Kittur et al. (2009) where workers collaboratively and incrementally refined a translation task output.

- **Gold standards**: CrowdFlower and Mechanical Turk support the ability to seed task units with pre-curated answers known as the ground truth. The ground truth is interspersed with the required task questions and task requesters can either monitor workers who deviate from the gold standard, or automatically discard their results. This potentially prevents submissions from malicious workers as well as surfacing possible issues with the task instructions.

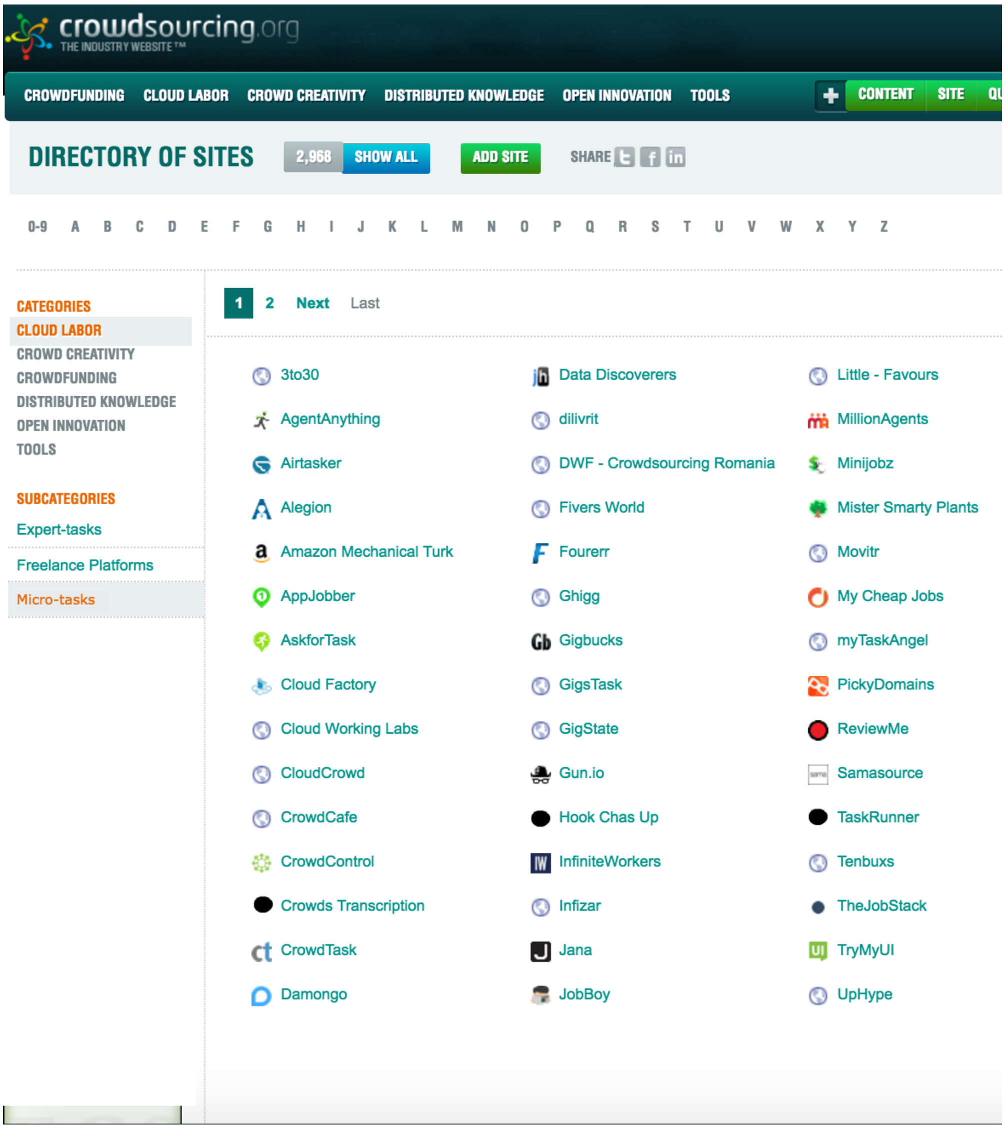
- **Expert review**: Mechanical Turk allows task requesters to review submissions from workers before they are accepted and before payment is made. This potentially deters malicious submissions which would not receive compensation. Unlike Mechanical Turk, CrowdFlower does not allow requesters to review tasks before payments are made, however, requesters can flag workers and remove them from their present and future tasks due to the quality of their submissions. The downside with expert review occurs when requesters flag workers who genuinely did not understand the task instruction. This has led researchers such as Irani and Silberman (2013) to build tools whereby workers can also review requesters who post tasks with unclear guidelines.

- **Reputation**: closely linked with expert reviews is the reputation system built around crowdsourcing platforms. Workers build their reputation by submitting high quality answers which in turn gives them access to higher paying tasks. Platforms such as Mechanical Turk and CrowdFlower employ a reputation system and offer requesters the ability to restrict their tasks to top performers. Requesters can also restrict their tasks by IP addresses thus limiting flagged workers from re-opening new accounts after stacking up a bad reputation.

2.4 Paid Microtask Crowdsourcing

Money is a natural incentive for carrying out work, hence paid crowdsourcing remains probably the most prominent form of crowdsourcing (Frei, 2009). Paid microtasks differ in not relying primarily on the goodwill, altruistic nature, love for science or fun factor that other crowdsourcing models tap into (e.g., citizen science projects). A survey of 50 paid crowdsourcing platforms was presented by Frei (2009) where they split up the systems along the lines of work type and work category. Complex projects were carried out on platforms such as InnoCentive, simple projects on eLance, macro tasks on LiveWork, and microtaks on Mechanical Turk. The platforms represented a thriving diverse economy, with over $1 billion paid out to over 1 million workers over a 10 year period (between 1999 and 2009). They also presented the volume to payment landscape, with open innovation platforms such as InnoCentive making very few payments valued in tens of thousands of dollars, while Mechanical Turk making millions of payments valued at less than a dollar. Our work however is primarily concerned with the microtask market place, currently dominated by Amazon's Mechanical Turk, but also featuring players such as CrowdFlower, which was used for all our tasks and experiments. Figure 2.7 presents a snapshot of microtask crowdsourcing platforms from `crowdsourcing.org`. The platform lists about 135 microtask sites (45 of which are displayed in the figure) out of a directory listing of about 3,000 sites.

Microtasks are deployed on highly parallelizable work pieces that can be solved at the micro level and re-aggregated to a useful piece for the requester. The classes of tasks amenable to paid microtask crowdsourcing generally involve: information finding; verification and validation; interpretation and analysis; content creation; surveys and content access presented earlier in Section 2.3.1. These are tasks that can be easily broken down into small pieces, and as such; requesters can deploy units at scale by leveraging on an increased budget and a large pool of available workers. This has led to a number of studies to understand the effect of financial incentives on task completion rates, volume of work done and quality of submissions. Research has shown that increased payment leads to faster task completion, howbeit, not at a higher quality (Mason and Watts, 2010). The dynamics of finding a balance between speed and quality has caused payouts on MTurk for example, to steadily rise from $0.10 in 2011 to about $0.50 in 2015 based on longitudinal research carried out by Difallah et al. (2015). A number of other factors however also play a role in obtaining task results at a high quality such as: bonuses,



3to30
AgentAnything
Airtasker
Alegion
Amazon Mechanical Turk
AppJobber
AskforTask
Cloud Factory
Cloud Working Labs
CloudCrowd
CrowdCafe
CrowdControl
Crowds Transcription
CrowdTask
Damongo
Data Discoverers
dilivrit
DWF - Crowdsourcing Romania
Fivers World
Fourerr
Ghigg
Gigbucks
GigsTask
GigState
Gun.io
Hook Chas Up
InfiniteWorkers
Infizar
Jana
JobBoy
Little - Favours
MillionAgents
Minijobz
Mister Smarty Plants
Movitr
My Cheap Jobs
myTaskAngel
PickyDomains
ReviewMe
Samasource
TaskRunner
Tenbuxs
TheJobStack
TryMyUI
UpHype

Figure 2.7: A snapshot from a crowdsourcing directory (crowdsourcing.org)

worker perception, and the variation of payment sizes across tasks (Mason and Watts, 2010). The drive to engineer optimal quality, speed and volume with minimal financial payments has made paid microtasks the subject of numerous discourses on the ethics of compensation (Irani and Silberman, 2013).

While monetary payments can be seen as an extrinsic motivator, other intrinsic factors have been known to contribute to sustained participation in crowdsourcing settings. Furthermore, certain tasks have been known to be more enjoyable than others: e.g., locating a celebrity Twitter handle vs. locating a journalist's Twitter handle; writing a review on an iPhone game vs. a review on a cisco router; and writing a 10 page paper on celebrity pets vs. a paper on a health reform bill (Frei, 2009). The question then was how do we expand the currency of transactions on paid microtask platforms to transcend monetary payments, and encompass tasks that have an intrinsic appeal. Framing tasks this way would be more attractive to workers, cost less for requesters, and potentially be solved quicker and at a higher quality. This led us to research into the gamification of paid microtasks.

Research area	Reference	Findings
Motivation & Incentives	Rogstadius et al. (2011)	intrinsic factors, such as framing a task as helping others – improves output quality where extrinsic motivators such as increased pay do not.
Gamification	Feyisetan et al. (2015b)	gamification leads to better accuracy and lower costs. It also makes paid microtask work more rewarding and engaging, with sociality features.
Collaboration	Kulkarni et al. (2012); Kittur (2010)	there is potential for gains in effort, motivation, coordination, and quality that can be achieved by letting people work together collaboratively.
Contests	Zheng et al. (2011)	intrinsic motivation is important than extrinsic in inducing participation. Autonomy and variety are associated with intrinsic motivation.
Cheating & Spamming	Difallah et al. (2012)	adversarial techniques include random answers, automated and semi automated answers, agreement on answers and answer sharing.
Quality control	Allahbakhsh et al. (2013)	quality control approaches include input and output agreement, consensus, expert review, ground truth and real-time support.
Demographics	Ross et al. (2010); Ipeirotis (2010b)	population has changed over time, from a primarily U.S. workforce to an increasingly international group of young, well-educated Indians.
Wages & Compensation	Horton and Chilton (2010)	the reservation wage – the smallest wage a worker is willing to accept for a task are approximately log normally distributed.
Legal & Ethics	Felstiner (2011)	why workers in particular subsections of the paid crowdsourcing industry may be denied the protection of employment laws.
Task routing	Bragg et al. (2014)	iterative methods for dynamically allocating batches of tasks that make near-optimal use of available workers in each round.
Workflows	Kittur et al. (2011, 2012)	by using coordination between workers, complex artifacts can be effectively produced by individuals contributing small amounts of time and effort.
Gold standards	Aroyo and Welty (2013)	perfect gold standards are a myth, there is not only one universally constant truth, disagreement should be embraced.
Performance	Mason and Watts (2010)	increased pay increase workers' willingness to accept a task or the speed of task completion, but do not improve the work quality.
Delivery speed	Bernstein et al. (2011)	With synchronous crowds, systems can dynamically adapt tasks by leveraging available workers who can be recruited within two seconds.
Future directions	Kittur et al. (2013)	research challenges in twelve major areas including workflows, real-time response, synchronous collaboration and motivation.

Table 2.2: A Few Research Areas in Paid Microtask Crowdsourcing

Table 2.2 presents a list of active research areas in paid microtask crowdsourcing. In the next few sub sections, we expand on the first four research areas (motivation, gamification, collaboration and contests) which are directly relevant to our approach to address challenges in crowdsourcing. Motivation in crowdsourcing had earlier been introduced in Section 2.3.3, while collaboration and contests were presented in Section 2.3.4 as approaches to carrying out crowdsourcing. These three topics are expanded in the following sections 2.4.1, 2.4.3 and 2.4.4. Collaboration serves as the basis for Chapter 9 on synchronous collaboration in crowdsourcing; while contests serve as the background on which Chapter 7 on real-time delivery is built. We also introduce the concept of gamification in crowdsourcing, which together with the earlier background on motivation, serves as background specifically for Chapter 8 and generally for much of the thesis. Some other research areas listed in Table 2.2 include quality control, cheating and spamming introduced earlier in Section 2.3.5 and demographics discussed in the 'who' of crowdsourcing in Section 2.3.2.

2.4.1 Motivation in crowdsourcing

Understanding human motivation has a long history of research in the social sciences and psychology. The Cognitive Evaluation Theory (CET) (Deci and Ryan, 1985a) described motivation to execute a task as being dependent on context factors of the task fulfilling basic psychological needs. The psychological needs highlighted were the need for autonomy, competence and social relatedness. This theory provides explanations for intrinsic motivations (Deci and Ryan, 1975) but misses some factors that cause people to engage in tasks. The Self Determination Theory (SDT) (Ryan and Deci, 2000) was proposed by the same authors as an extension to the Cognitive Evaluation Theory. Whereas CET focused on intrinsic motivation SDT extended to extrinsic motivation which serves on a continuum of bringing the individual to intrinsic self-motivation. Other theories from psychology include the General Interest Theory (GIT) (Eisenberger et al., 1999) which emphasises the relevance of the task as the core motivator to its successful performance. GIT also presented a breakaway from CET/SDT, which posited that external incentives and rewards hampers intrinsic autonomy, hence negatively affecting performance. GIT argues that rewards can positively affect performance by increasing intrinsic motivation, however, financial rewards act more as a two-edged sword – having either positive or negative consequences. Positive when rewards affirm competence; and negative when individuals are not clear on how to attain the reward i.e., the reward criteria is vague. This leads into the work by Kerr (1975) which stresses the importance of rewarding individuals for clearly defined and expected behavioural outcomes. Kerr (1975) presented a management perspective to motivation and rewards with examples from politics, business and medicine, illustrating how rewards should be aligned with desired behaviour. Table 2.3 and 2.4 presents constructs of intrinsic and extrinsic motivations from a study by Kaufmann et al. (2011).

Within the context of community driven circles, several researchers have sought to understand why individuals take part in certain tasks. The obvious answer is – *they are paid to do it*. This has led to a number of studies on the role of financial rewards in incentivising workers such as Mason and Watts (2010); Rogstadius et al. (2011); Horton and Chilton (2010); Yin et al. (2013); Harris (2011) and Ho et al. (2015). However, according to research such as by Kaufmann et al. (2011), money is but a piece of the puzzle. For example, Wikipedia is currently the worlds largest encyclopedia – built up from the

Construct	Example
Skill variety	A worker picks a translation task because he likes translating and wants to use his skills in his favorite foreign language.
Task identity	A worker picks a task because it allows him to see how the result of his work will be used – e.g. writing a product description for a website
Task autonomy	A worker who is motivated because a certain task allows him to be creative – e.g. designing a logo or a website.
Job direct feedback	A worker who is motivated because a task provides the opportunity to check if his result is correct – e.g. a programming task.
Past-time	A worker who uses the platform or works on various random tasks because he has nothing better to do.
Community identification	A worker, who only accepts tasks from requesters with a good reputation, because they are known as valuable supporters of the platform and its community.
Social contact	A person is active on a crowdsourcing platform just to meet new people

Table 2.3: Constructs of Intrinsic Motivation Kaufmann et al. (2011)

ground by unpaid contributing individuals. The motivation of Wikipedia contributors have also been studied extensively in a bid to understand, and possibly replicate its success (Kuznetsov, 2006; Nov, 2007; Schroer and Hertel, 2009; Rafaeli and Ariel, 2008). However, with these studies, it has been difficult to re-engineer a Wikipedia-styled success story. Similar to Wikipedia would be understanding the motivations of software developers that contribute numerous hours into open source software – notably the Linux OS and Apache projects. Several motivators have been suggested such as altruism, community identification, enhanced status, intellectual stimulation, future rewards, learning and personal beliefs (Lakhani and Wolf, 2005; Hertel et al., 2003; Roberts et al., 2006; Hars and Ou, 2001; Ye and Kishida, 2003; Bitzer et al., 2007). Individuals also voluntarily assume roles (devoid of financial compensation), including project leaders, core members, active developers, peripheral developers, bug fixers, bug reporters, readers and passive users – in decreasing order of involvement (Ye and Kishida, 2003). These show the variegated nature of what might motivate and incentivise participation in tasks.

Table 2.5 from Smart et al. (2014) presents dimensions and corresponding characteristic values in defining motivation among crowd workers. Apart from the intrinsic and extrinsic motivation types early decribed by Kaufmann et al. (2011), Smart et al. (2014) also lists different forms of motivation which are applicable to crowdsourcing. These include economic forms (which are the most common incentives used in paid microtasks); altruistic incentives (such as disaster relief tasks or malaria test annotations); hedonic (which

Construct	Example
Payment	A worker is active on a crowdsourcing platform as a form of primary or secondary income.
Signaling	A worker who joins a platform or selects tasks in order to show presence and advance his chance of being noticed by possible employers.
Human capital advancement	A worker picks translation tasks because he or she wants to improve language skills for a new or better job.
Action significance by external values	A worker joins a platform and participates because the values it stands for are important to him as well (e.g. freedom of speech).
Action significance by external obligations	A student working on scientific survey tasks on a crowdsourcing platform because he is obliged to do so by his professor / tutor.
Job indirect feedback	A worker is very committed because he seeks commendation.

Table 2.4: Constructs of Extrinsic Motivation by Kaufmann et al. (2011)

Dimension	Values / Characteristics	Example
Motivation type	Intrinsic	Bitzer et al. (2007)
	Extrinsic	Rogstadius et al. (2011)
Form of motivation	Economic	Mason and Watts (2010)
	Altruistic	Mavandadi et al. (2012)
	Hedonic	von Ahn and Dabbish (2004)
	Reputational	Archak (2010)
	Instrumental	Brabham (2013)
	Other	Howe (2006)
Reward type	None	Poesio et al. (2015)
	Monetary payment	Horton and Chilton (2010)
	Prize	Rokicki et al. (2014)
	Other	Howe (2006)
Reward variability	Fixed	Kaufmann et al. (2011)
	Variable	Mao et al. (2013a)
	None	Mason and Watts (2010)

Table 2.5: Dimensions of motivation and incentives by Smart et al. (2014)

features in gamified systems – more details in the next section); reputational incentives (which were described earlier in Section 2.3.5 as a form of quality control); instrumental and all other forms. The reward types could be none (for example in game based crowdsourcing systems such as Phrase Detectives by (Poesio et al., 2015); monetary payment; or prize based such as in crowdsourcing contests such as TopCoder (Archak, 2010). The rewards could also be variable based on the volume of work done or fixed.

2.4.2 Gamification in Crowdsourcing

Gamification is the use of game design elements in non-game contexts in order to achieve the effects of fun and engagement that derives from playing a game (Zichermann, 2011). This include systems that build a complete game narrative around a task (e.g., FoldIt[9] and EyeWire[10]); those that employ tactics such as micro-diversions to fend off boredom; tasks designed to further a noble cause or stoke curiosity; and systems that engineer game elements into tasks. The idea of using gameful tactics to spur productivity is not a new idea. According to Nelson (2012), in the early to mid 20th century, the Soviet Union created games to increase productivity, via experiments ranging from purely competitive games directly tied to productivity, to attempts at morale-building via team games and workplace self-expression. Badges have also been handed out as a symbol of achievement in the Boy Scouts of America since the early 20th century (Deterding, 2012). McGonigal (2011) traces the utility of games even further down an earlier time in history where alternating one day of playing games and one day of eating, sustained a nation through eighteen years of farming. McGonigal (2011) also estimates that there are over 5 million ‘extreme’ gamers in the US playing over 45 hours of games every week – detached from the reality of the world and immersed in the virtual world of games. If games have such a great appeal, then the desire to harness the potential power gameplay becomes apparent: either convert current gamer output into useful work, or design entire tasks around a game narrative, or adopt the game elements that afford for engagement into non gaming contexts.

Table 2.6 presents a hierarchical abstraction of layers in gamification. It starts with the abstract game design methods such as play-centric design and value conscious design; to various game models such as challenge, curiosity and fantasy games. The next layer lists various principles, followed by design patterns and game mechanics. Zichermann (2011) lists twelve mechanics. The final layer consists of the actual game elements visible on the interface. A run through of game elements by Seaborn and Fels (2015) is listed in Table 2.7. The twelve game design patterns and mechanics by Zichermann (2011) include:

- Pattern recognition
- Collecting (e.g. badges)
- Surprise (e.g. easter eggs)
- Organising (e.g. time challenge)
- Gifting (e.g. karma points)
- Flirtation (e.g. poking)
- Recognition (e.g. trophies)
- Leading others (e.g. teams)
- Fame (e.g. leaderboards)
- Being the hero (e.g. missions)
- Status (e.g. public badges)
- Nurturing (e.g. tamagotchi)

[9] Foldit – https://fold.it/
[10] EyeWire – http://eyewire.org/

Level	Description	Examples
Game interface design patterns	Common, successful interaction design components and design solutions for a known problem in a context, including prototypical implementations	Badge, leaderboard, level curiosity.
Game design patterns and mechanics	Commonly reoccurring parts of the design of a game that concern gameplay	Time constraint, limited resources, turns.
Game design principles and heuristics	Evaluative guidelines to approach a design problem or analyze a given design solution	Enduring play, clear goals, variety of game styles.
Game models	Conceptual models of the components of games or game experience	Mechanics Dynamics Aesthetics (MDA); challenge, fantasy, curiosity; game design atoms; Core Elements of the Gaming Experience (CEGE).
Game design methods	Game design-specific practices and processes	Playtesting, playcentric design, value conscious game design.

Table 2.6: Levels of game design elements by Deterding et al. (2011a)

Term	Definition	Alternative
Points	Numerical units indicating progress	Experience points; score.
Badges	Visual icons signifying achievements.	Trophies.
Leaderboards	Display of ranks for comparison.	Rankings, scoreboard.
Progression	Milestones indicating progress.	Levelling, level up.
Status	Textual monikers indicating progress.	Title, ranks.
Levels	Increasingly difficult environments.	Stage, area, world.
Rewards	Tangible, desirable items.	Incentives, prizes, gifts.
Roles	Role-playing elements of character.	Class, character.

Table 2.7: Game element terminology by Seaborn and Fels (2015)

Surveys such as by Hamari et al. (2014) and Seaborn and Fels (2015) paint a picture of current research trends in gamification. The most popular gamification elements used are points, leaderboards, badges and levels. Table 2.7 gives a definition of these elements. However, beyond these game elements, gamification features dynamic mechanisms that foster engagement and motivation. Blohm and Leimeister (2013) stated six of such mechanisms, three of which include *collection*, *collaboration* and *competition* which were earlier discussed in Section 2.3.4 on the 'how' dimension of crowdsourcing. The gamification mechanisms include:

1. exploration, which motivates intellectual curiosity;
2. collection of badges and trophies, which brings a sense of achievement;
3. acquisition of status, which makes individuals strive for social recognition;
4. collaborative group tasks, which facilitate social exchanges;
5. time pressure challenges, which engender cognitive stimulation; and
6. organization in virtual worlds that creates a self-determination desire.

Gamification done right is beyond 'pointsification' (merely tacking on points, badges and leaderboards) and is able to withstand scrutiny by notable anti-gamification critics such as Bogost (2011, 2015).

Game mechanics	Game dynamics	Motives
Documentation of behaviour	Exploration	Intellectual curiosity.
Scoring systems, badges, trophies	Collection	Achievement.
Rankings	Competition	Social recognition.
Ranks, levels, reputation points	Acquisition of status	Social recognition.
Group tasks	Collaboration	Social exchanges.
Time pressure, tasks, quests	Challenge	Cognitive stimulation.
Avatars, virtual worlds	Organization	Self-determination.

Table 2.8: Game design elements and motives by Blohm and Leimeister (2013)

Gamification practices have also raised questions about its potential to undermine innate intrinsic motivation, known as replacement and over-justification (Zichermann, 2011). For example, a child that naturally loves to play the violin might altogether lose the desire to play on the introduction and removal of a competitive reward system (Frey and Jegen, 2001). Other negative outcomes that need to be paid attention to include the effects of increased competition, task evaluation difficulties, and understanding design features (Hamari et al., 2014). These downsides have been shown to be task dependent and can thus be mitigated. The analysis from Hamari et al. (2014) and Seaborn and Fels (2015) show that majority of the experiments in literature had positive results in terms of: response speed, quality, enjoyment, learning, compliance, satisfaction, collaboration, participation and engagement on the introduction of gamification.

2.4.3 Collaboration in Crowdsourcing

Paid microtask crowdsourcing has traditionally been approached as an individualistic endeavour. Individual workers complete tasks without interacting with others. Even most interdependent crowdsourcing tasks, such as employ the 'find-fix-verify' workflow pattern, adopt a serial synchronous approach (Bernstein et al., 2010). This can still pass as a form of collaboration, albeit, not in a real-time interactive fashion. For example, in a paragraph-shortening task, one worker might identify an area that can be shortened without changing the meaning of the paragraph; another edits the highlighted section to shorten its length; while a final worker verifies the edit. Other platforms such as CrowdForge by Kittur et al. (2011) also apply a map-reduce approach to splitting up tasks among workers and re-aggregating a single result. Throughout these workflows, the workers do not interact directly, despite relying on the output of each other to kick-start their own sub-task.

On the hand, other forms of crowdsourcing, such as citizen science projects, have embraced richer models that feature increased collaboration and interaction between participants to great success (Tinati et al., 2015). This is beneficial in complex tasks that might require self-organization, idea sharing and discussions. It is also more representative of what obtains in the real world with office workers and academic researchers interacting and collaborating to achieve a common goal.

Other researchers have attempted other collaboration strategies drawing inspiration from existing research (Greenberg and Bohnet, 1991). For example, Lasecki et al. (2012b) demonstrated that crowd workers have the ability to retain knowledge after a task has ended, and pass on the knowledge to new workers on the task (in a fashion akin to organisational learning). Rokicki et al. (2015) presented mechanisms for team based crowdsourcing competitions. In some instances, workers were permitted to self-select a team – usually joining the leading team or merging smaller teams to challenge the leading team. Their work also highlighted the importance of team communication, which was crucial to admitting new members, discussing the overall approach to the task and clarification of rules.

Anagnostopoulos et al. (2012) also studied team formation dynamics for solving tasks, however, they designed algorithmic approaches for assigning team members based on their individual skills. They reported this as having the potential to improve team coordination and collaboration. Kittur (2010) designed a translation platform which sourced workers from Mechanical Turk. The workers autonomously collaborated and coordinated with each other to translate and refine the original text, yielding a result that was voted better that a professional translation.

2.4.4 Contests in Crowdsourcing

Revisiting the image in Figure 2.1, we observe that one approach to crowdsourcing is via contests or competitions. Contests is also one of the three methods we discussed in the ‘how’ of crowdsourcing in Section 2.3.4 (alongside collection and collaboration discussed above). The innovation prizes by Netflix and X-Prize; and competition platforms such as TopCoder Archak (2010) serve as a medium to elicit the single best response from a crowd of participants. One of the best-known crowdsourcing contests in social computing and crowdsourcing took place in 2009. DARPA set up a challenge to locate 10 red weather balloons within the continental United States (Tang et al., 2011). The winning team from MIT employed crowdsourcing strategies to leverage on a multi-level network of people and their friends. Crowdsourcing contests are traditionally deployed when the requester seeks one best or final answer (as opposed to an aggregation of worker results). For example, the Netflix $1million challenge to build a better recommendation algorithm (Bennett and Lanning, 2007) or the $10million Ansari X Prize [11] both fall in this category of best response. However, as mentioned in the beginning of this chapter, taking advantage of the crowd via inducement prizes dates back to the 18th century. Some of the most notable contests in history [12] are presented in Table 2.9

Prize	Definition	Year
British Longitude prize	Determination of a ship's longitude at sea	1714
The Alkali Prize	Method to produce alkali from sea salt	1775
Food Preservation Prize	Preserving food on long military campaigns	1795
Turbine Prize	Commercially viable hydraulic turbine	1823
The Rainhill Trials	Railway locomotives	1829
Substitute for Guano Prize	Alternative to the guano manure	1852
The Billiard Ball Prize	Alternative material to elephant ivory	1863
Butter Substitute Prize	A cheaper substitute for butter	1869
The Schneider Cup	Seaplanes and flying boats	1913
The Orteig Prize	Non-stop flight from New York to Paris	1919

Table 2.9: Historical Challenge Prizes

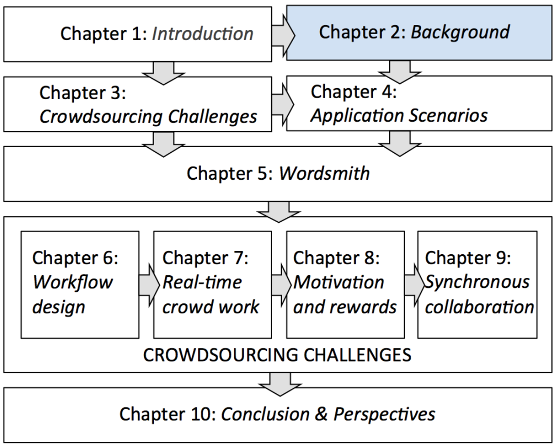
Remuneration in contests range from the winner-takes-all scenario, which compensates only the best participant; to more relaxed models that pay contributors who make submissions above a certain threshold. Contests also leverage on the urgency that comes from a fixed time frame; and the satisfaction that a sense of winning brings. In this context, one of the most popular crowdsourcing contest platform is TopCoder. TopCoder hosts weekly algorithms and software design competitions where winners receive financial remuneration and performance points (Archak, 2010). The contests are time constrained with the algorithm competitions lasting 2 hours, and the system design lasting one week. The software generated is licensed for profit to companies while the contestant are paid and rated. The rating system is of particular importance because it serves as a recruitment platform for companies to access the best developer talents. One of the outcomes of the study by Archak (2010) was that the online reputation of individuals in crowdsourcing contests have significant economic value. Winning for contestants went beyond the immediate financial payout. Submitting a good enough result beyond a threshold was enough to boost ratings that would yield future economic dividends.

[11] http://ansari.xprize.org/

[12] Nesta-http://www.nesta.org.uk/news/guide-historical-challenge-prizes

As with gamification and collaboration, contests indicate that the motivation and desire to participate in crowdsourcing tasks transcend monetary compensation. Indeed, studies and interviews with crowd workers by Felstiner (2011) and Martin et al. (2014) have shown that many workers perform crowdsourcing tasks to get paid. However, with gamification, collaboration and contests, the task becomes slightly different. An element of engagement and enjoyment is introduced which can serve as an additional currency of transaction in paid microtask crowdsourcing.

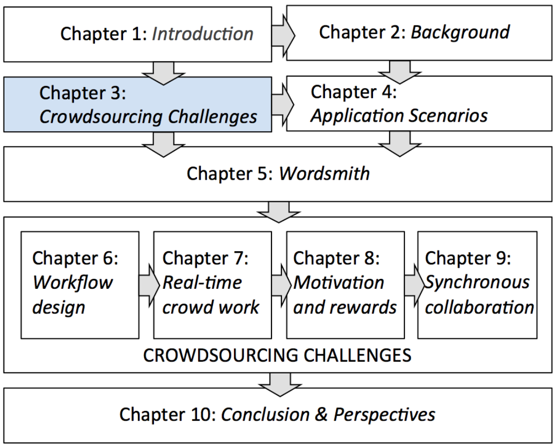
2.5 Summary


Chapter 1: *Introduction*
Chapter 2: *Background*
Chapter 3: *Crowdsourcing Challenges*
Chapter 4: *Application Scenarios*
Chapter 5: *Wordsmith*
Chapter 6: *Workflow design*
Chapter 7: *Real-time crowd work*
Chapter 8: *Motivation and rewards*
Chapter 9: *Synchronous collaboration*
CROWDSOURCING CHALLENGES
Chapter 10: *Conclusion & Perspectives*

In this chapter we presented background material that formed the foundations of the thesis. We introduced related socio-technical fields that helped us better understand the role of crowdsourcing. We also took an in-depth look at the state of the art in crowdsourcing with specific emphasis on paid microtask crowdsourcing. The last part of the chapter gave an overview on motivating paid microtask workers with money and additional incentive mechanisms. We discussed gamification, collaboration and contests as three broad categories of incentives on which we base future chapters when addressing specific challenges in the thesis.

Chapter 3

Crowdsourcing Challenges


Chapter 1: *Introduction*
Chapter 2: *Background*
Chapter 3: *Crowdsourcing Challenges*
Chapter 4: *Application Scenarios*
Chapter 5: *Wordsmith*
Chapter 6: *Workflow design*
Chapter 7: *Real-time crowd work*
Chapter 8: *Motivation and rewards*
Chapter 9: *Synchronous collaboration*
CROWDSOURCING CHALLENGES
Chapter 10: *Conclusion & Perspectives*

This chapter gives an in-depth outlook on the research challenges studied in the thesis. We discuss issues in workflow design; real-time crowd work; motivation and incentives engineering; and collaboration in paid microtask crowdsourcing. We present introductory background material on each challenge, discuss the state of the art in current research and then outline the specific parts of the challenge that this work seeks to address.

3.1 Workflow Design

Harnessing the rapid increase in the generation of data has led to advances in the World Wide Web, the Semantic Web and the Web of Data (Auer et al., 2007) – translating into the need to crowdsource useful information to fulfil their visions. A first step in making sense of the data necessitates information extraction and annotation of datasets. This has led to the availability of training datasets for Natural Language Processing algorithms from research such as ACE (Doddington et al., 2004), MUC (Chinchor, 1998) and CoNLL (Tjong et al., 2003). An important task in this context is the identification of named entities - the people, places, organisations, and dates referred to in text documents - and their mapping to Linked Data URIs (Usbeck et al., 2014). This importance therefore informed our task selection to address the issue of workflow design in paid microtask crowdsourcing. State-of-the-art technology in entity recognition achieves

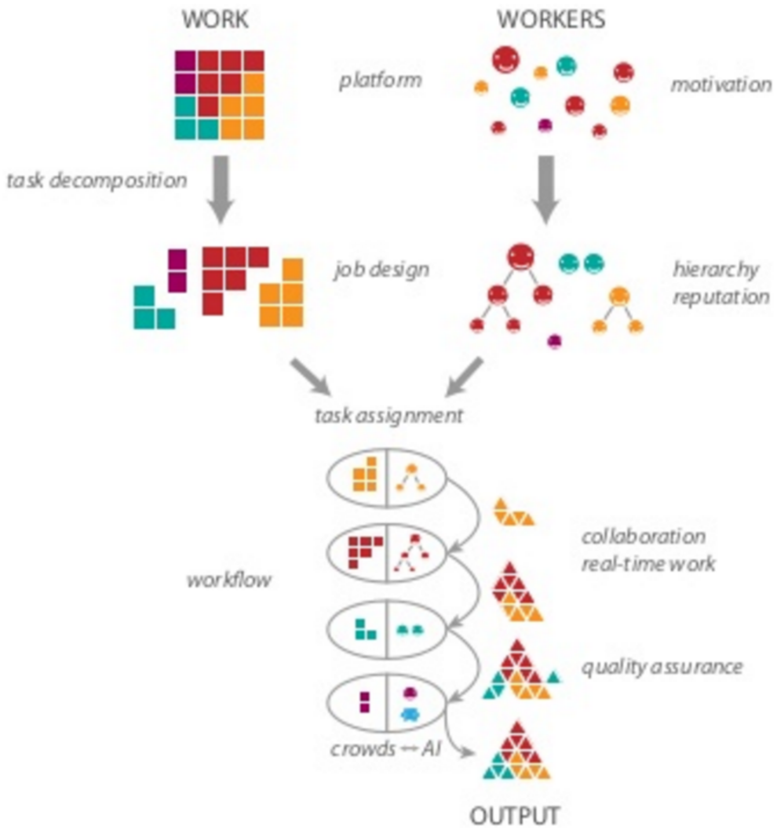


Figure 3.1: Future Model of Crowd Work by Kittur et al. (2013)

near-human performance for many types of unstructured sources; most impressively so for well-formed, closed-domain documents such as news articles or scientific publications written in English (Marrero et al., 2009; Nadeau and Sekine, 2007). It has been less successful so far in processing social media content such as microblogs, known for its compact, idiosyncratic style (Derczynski et al., 2015). Human computation and crowdsourcing offer an effective way to tackle these limitations (Snow et al., 2008), alongside increasingly sophisticated algorithms capitalising on the availability of huge data samples and open knowledge bases such as DBpedia and Freebase (Rizzo and Troncy, 2011).

Advances in natural language processing have led to an understanding of textual structure which can be easily processed by computers (e.g., well formed news-wire articles with sufficient disambiguation context). Essentially, hybrid workflows have therefore led to pipelines which first selects text for machine annotation, passing the residue to the crowd (such as the approach by Demartini et al. (2012). These hybrid approaches to NER (named entity recognition) (Derczynski et al., 2015) that seamlessly bring together human and computational intelligence are however far from being the norm. While the technology to define and deploy them is on its way - for instance, tools such as GATE already offer built-in human computation capabilities (Sabou et al., 2014; Bontcheva et al., 2014a) and CrowdDB attempts crowd powered query engines (Trushkowsky et al., 2013) – little is known about the overall performance of machine-crowd-expert NER workflows and the factors that affect them. Besides various experiments reporting on task design, spam detection, and quality assurance aspects e.g., (Difallah et al., 2012; Snow et al., 2008; Yuen et al., 2011), at the moment we can only guess what features of a micropost, crowd contributor, or microtask platform will have an impact on the success of crowdsourced NER. The situation is comparable to the early stages of information extraction; once the strengths and weaknesses of particular methods and techniques had been extensively studied and understood, the research could then focus on overcoming real issues, propose principled approaches, and significantly advance the state of the art.

Workflows in paid microtask crowdsourcing began as simple parallelised tasks. A large piece of work was split up among multiple workers who solved individual task pieces. The need to solve more complex tasks led to more ingenious ways to split the request. Some approaches include: simple serial pipelines where the output of one worker is passed on to the next (Little et al., 2010); the *find-fix-verify* workflow where one worker identifies a task target, another set of workers carry out the task and a final worker verifies (Bernstein et al., 2010); and the various workflows inspired by the *map-reduce* framework of traditional computing (Dean and Ghemawat, 2008) where parallelised tasks are aggregated in stages until the final output is resolved (Little et al., 2010; Bernstein et al., 2010; Kittur et al., 2011).

3.1.1 Existing Workflows

One challenge in workflow design is facilitating coordination within the distributed workforce. This has been studied by a number of researchers like Kittur et al. (2008, 2009) some of whom have applied traditional techniques from computing such as van Der Aalst et al. (2003) and organisational literature such as Stohr and Zhao (2001). Other techniques that can constitute a form of workflow design include crowdsourcing contests (Cavallo and Jain, 2012; Dechenaux et al., 2014) or adopting some form of collaboration (Kittur, 2010) – although these would be addressed in details in future chapters.

- **CrowdForge** (Kittur et al., 2011): represents a framework and toolkit for crowdsourcing complex work. The CrowdForge approach is a simplified distributed computing methodology based on MapReduce introduced by Dean and Ghemawat (2008). The process consists of three subtasks: the first where a large task is partitioned into subtasks; the second phase where each subtask is 'mapped' or assigned to a worker to be solved; and a final 'reduce' phase where the outputs of workers are merged into a single final output. The process can be used to solve complex workflows by starting out with a root partition and recursively creating subtasks, mapping to workers and reducing intermediate stages. CrowdForge is effective when a task can be easily broken down into unit subtasks that can be solved independently – for example, a case study of writing an article was presented where the unit partitions involved topic sections and paragraphs.

- **CrowdWeaver** (Kittur et al., 2012): is a system built on CrowdFlower which supports visually managing complex crowd workflows by the management and reuse of templates. Task requesters can string tasks together, connected via dataflows between them. It also allows for monitoring and alerting based on worker task progress. Requesters are also able to split and merge tasks in a fashion similar to the previously discussed CrowdForge by Kittur et al. (2011).

- **TurkIt** (Little et al., 2010): is a set of APIs that allow task requesters to carry out iterative tasks to workers on Mechanical Turk. Rather than solving tasks in parallel, or using a MapReduce methodology, Turkit adopts an imperative approach where workers act as successive subroutines that change the state of the task until a final output is produced. An example is a task to decipher ineligible handwriting wherein workers iteratively solve the part they can with subsequent workers improving on earlier submissions.

- **Turkomatic** (Kulkarni et al., 2012): is a system for crowdsourcing complex jobs. Unlike CrowdForge and CrowdWeaver, Turkomatic engages the crowd in decomposing the jobs into multiple tasks which is then solved by multiple workers. The task requester can monitor the task decomposition process and intervene to improve the entire workflow. It employs a technique similar to the MapReduce approach from CrowdForge known as *price-divide-solve* (PDS). With PDS, workers recursively break down tasks (similar to the multiple partition steps of CrowdForge), then the tasks are solved and combined (similar to the reduce stage in CrowdForge). Turkomatic includes visualization and editing capabilities for requesters – similar to CrowdWeaver but absent from CrowdForge.

- **Jabberwocky** (Ahmad et al., 2011): is a full fledged crowdsourcing workflow framework also built on the MapReduce paradigm. It consists of a human and machine resource management system, a parallel programming framework based on MapReduce, and a high-level programming language. Unlike CrowdForge where the requester defines the partition steps, and Turkomatic where workers decide the divisions, decomposition in Jabberwocky can be either human powered or automatic.

3.1.2 Challenge

The specific workflow challenge we seek to address involve the design of hybrid systems for solving crowdsourcing tasks. One use case for crowdsourcing is using humans to

		CrowdForge	**CrowdWeaver**	**TurkIt**	**Turkomatic**	**Jabberwocky**
Definition language	paradigm	configuration	imperative, visual	imperative, textual	declarative	imperative, textual
	notation	wizard, python for custom procs	custom modeling language, visual	JavaScript like	-	Dog language
Task support	crowd platform provider	MTurk	CrowdFlower	MTurk	MTurk	self
	crowd management	–	–	–	–	profile-based preselection
	machine tasks definition	–	generic machine task	script	–	script
Control flow support	task instantiation	✓	✓	✓	✓	✓
	sequential instantiation	✓	✓	✓	✓	✓
	parallel execution	✓	✓	–	✓	✓
	via decision points	–	–	✓	–	✓
	looping / iterative execution	–	–	✓	–	✓
	crowd subprocess	✓	–	✓	–	✓
Data management support	data hosting type	data	data	data	data	data
	data passing among tasks	by value	data flow	by value	self-managed data flow	by value
	data splitting, aggregating	by crowd	built-in	script	by crowd	script
Development support	task design support	manual	wizard	manual	pre-defined	manual
	task deployment	automatic	automatic	automatic	automatic	automatic
Quality control support		voting	control questions, consensus	voting	voting	–
Public availability		open source availability	–	open source availability	–	–

Table 3.1: Analysis of crowdsourcing workflows by Kucherbaev et al. (2016)

perform tasks that computers cannot yet perform well e.g., image recognition. The data generated by the crowd is then fed back to computers to train them in carrying out the task, which they gradually get better at undertaking. Consequently after the training phase, a workflow scenario would attempt to solve a task first with the computer, passing along the more difficult pieces to the crowd. However, given that the general crowd is modelled as a homogeneous set of inexpert workers, it might become essential to further assign more difficult cases to a team of experts - an idea that has been termed *nichesourcing* an expert crowd by De Boer et al. (2012). Agreement between the crowd hence serves as an identifier for simple tasks which yields high consensus; disagreement on the other hand then serves as a signal source on the more difficult case that would require expert adjudicators as suggested by Aroyo and Welty (2013).

What we seek to tackle therefore, are methods to identify the tasks that sit in the middle of the workflow i.e., can we identify task features that make them ideal to be solved by the general crowd? Apart from the task features, we also seek insight into the behaviour of the crowd in the presence of task choice i.e., given the opportunity to skip through a task in lieu of confidence, what sort of sub-tasks would they be inclined to choose? Observing the behaviour of the crowd, and the features of the task they select, with respect to how well they perform the task, would help us decide what tasks they are best suited for.

In chapter 6 we begin addressing this first selected challenge in paid microtask crowdsourcing. We show that even seemingly simple tasks require a degree of intelligent design in getting the best out of the crowd. We demonstrate the challenge of designing workflows to carry out named entity recognition via crowdsourcing. On the surface, crowd based named entity recognition may seem like a trivial task, one that can be solved using a simple worklow as shown in the traditional crowdsourcing model in Figure 3.2. However, our experiments reveal this not to be the case. In fact, a 'black box' analysis (without checking the analysis for constituent precision and recall scores on specific entity types) reveals almost comparable accuracy scores between automatic entity recognition software (Derczynski et al., 2015) and crowdsourced entity recognition (Feyisetan et al., 2015a).

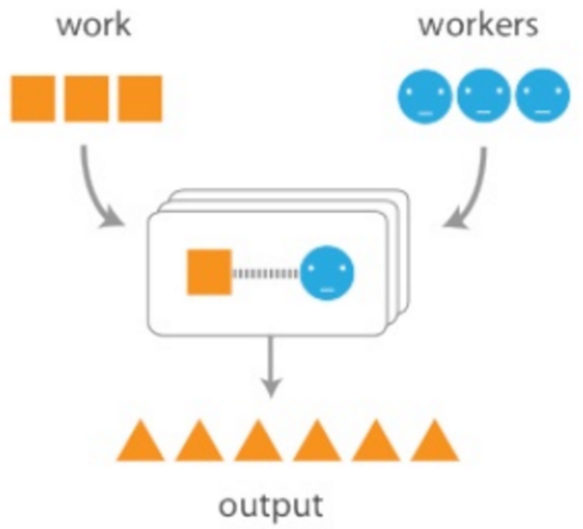


Figure 3.2: Traditional Crowdsourcing Workflow by Kittur et al. (2013)

A closer analysis of crowdsourced entities however suggests that although the crowd might be good at performing a task, they might be better at carrying out specific subtasks. Therefore, task decomposition needs to go beyond simplistic parallelisation of an entire task-set to available workers; and hybrid workflows combining machines and humans are thus essential to obtain top-percentile results. This approach (of using hybrid workflows) has been studied by researchers such as Demartini et al. (2012) who suggested a pipeline with machines carrying out a request and humans picking up tasks with low confidence results from the machines. However, within the difficult cases passed to the crowd, it is not unusual to encounter requests that defy traditional agreement and quality score metrics in terms of the final worker output. We posit that this might be as a result of one or two things: (i) the 'byte' sized nature of microtasks attunes the crowd to carry out some sub-tasks much easily and with lower cognitive overhead than other sub-tasks within the same task; and/or (ii) some sub-tasks might be genuinely difficult, with the potential to raise conflicting answers between crowd workers. The challenge then lies in decomposing the task using insights gained from the task type and worker interactions, as opposed to adopting a generic split approach.

Hybrid annotation techniques have emerged as a promising approach to carry out named entity recognition on noisy microposts. In chapter 6, we identify a set of content and crowdsourcing-related features (number and type of entities in a post, average length and sentiment of tweets, composition of skipped tweets, average time spent to complete the tasks, and interaction with the user interface) and analyse their impact on correct and incorrect human annotations. We then carry out further studies on the impact of extended annotation instructions and disambiguation guidelines on the factors listed above. These are all done using CrowdFlower and our bespoke crowdsourcing platform (introduced in chapter 5) on three datasets from related literature and a fourth newly annotated corpus. Our findings show that crowd workers correctly annotate shorter tweets with fewer entities, while they skip (or wrongly annotate) longer tweets with more entities. Workers are also adept at recognising people and locations, while they have difficulties in identifying organisations and miscellaneous entities, which they skip (or wrongly annotate). Finally, detailed guidelines do not necessarily lead to improved annotation quality. These findings lead to the design of more advanced NER pipelines, informing the way in which tweets are chosen to be outsourced to automatic tools, crowdsourced workers and nichesourced experts.

3.2 Real-time Crowd Work

An increasing number of application scenarios require microtask platforms to deliver responses in near real-time. These tasks come with hard deadlines or tight constraints which diminish the value of result outputs the later they arrive. Longitudinal studies of crowdsourcing marketplaces such as Amazon's Mechanical Turk (Difallah et al., 2015) reveal how microtasks have transitioned from outsourcing platforms where work took days to complete (Kittur et al., 2008), down to those which took hours to complete (Ipeirotis, 2010a), finally achieving delivery times in the range of seconds and under (Bernstein et al., 2011). Surveys such as Mason and Watts (2010) have shown that increasing the payoff of microtasks results in tasks being completed faster; however, this does not guarantee the response times required in time-critical scenarios. For near real-time behaviour to be a reality in microtask crowdsourcing, two components are critical:

1. **Timely worker recruitment** – pulling together large *flash crowds* that arrive within moments of the task; and

2. **Timely task completion** – collective completion of tasks by workers efficiently within the required constraints.

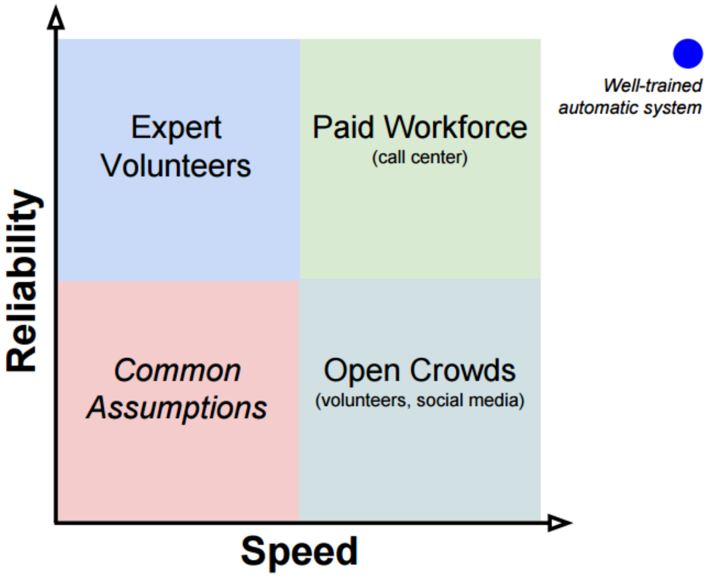


FIGURE 3.3: Speed and reliability trade-offs in crowdsourcing by Lasecki et al. (2014)

The real-time speed benefits sometimes come at a cost of accuracy and reliability as highlighted by Lasecki et al. (2014), and shown in Figure 3.3 with the ultimate desire to achieve results close to automatic systems (the blue dot in the figure).

3.2.1 Timely Worker Recruitment

Several models have been proposed to ensure timely availability of crowd workers. With Adrenaline, Bernstein et al. (2011) employed a retainer model to have access to flash crowds on demand. Workers were paid a retainer fee - between $0.05 to $0.10 per minute – to be on call in order to respond promptly to a new task. During the wait time, they are free to take on other tasks. When a new task becomes available, workers are alerted via an audio chime. They reported response rates within two seconds, with 75% joining the task in three seconds. Some other researchers have experimented with sending emails ahead of time, specifying task start times (Kittur et al., 2013). In Viz-Wiz, Bigham et al. (2010) used the quikTurKit approach to recruit workers in advance. When quikTurkit detected that a new task might be soon available, it starts recruiting new workers. The workers are kept busy on previous tasks until the new task arrives. Viz-Wiz also employs a simple but effective approach of posting tasks repeatedly on the marketplace platform. Multiple posts ensure that the task remains on the first page of the platform to ensure prospective workers have it in their view. Other approaches relied on recruiting a much larger crowd than required. Lasecki et al. (2014) adopted this approach of larger crowds on the premise that the fastest workers can be recruited - which further speeds up the task. Bernstein et al. (2012) used queuing theory with predictive recruitment (precruitment) to speed up the worker acquisition process. Precruitment is similar to quikTurkit and the retainer model in that it attempts to contact workers before tasks arrive. Precruitment however leverages on the retainer model to further drop response times to 500ms.

3.2.2 Timely Task Completion

However, mobilizing workers does not automatically translate into timely completion of tasks (Bernstein et al., 2011). Some methods have been presented in literature to address timely completion. Bernstein et al. (2011) designed Adrenaline as a smart camera shutter which captures a short video and used the crowd to instantaneously select the best photo frame. In Adrenaline, *rapid refinement* is used to cut down crowd latency by dynamically narrowing the search space following early signs of worker agreement. It speeds up the task by focusing slower workers on a smaller search space. *Stream parallelism* is applied on continuous tasks by dynamically assigning roles to workers (Lasecki et al., 2014). For example, stream parallelism is used in WeGame by Loparev et al. (2014) which merges streams of multiple players controlling a single game character in real-time. A form of stream parallelism is also utilised in Viz-Wiz (Bigham et al., 2010) and Legion (Lasecki et al., 2011). Most real-time tasks cannot be addressed within the time frame of its arrival – for example, transcribing or translating a live speech. With *temporal division*, the streaming task is divided into small manageable segments (e.g. a single sentence to transcribe) across the workers. This approach is used in tandem with *stream parallelism* to provide for multiple redundant task performance (Lasecki et al., 2014). This technique can be seen in Scribe (Lasecki et al., 2012a) which provides real-time captioning of conversations and live events to deaf people. It is also utilised in Legion:AR (Lasecki et al., 2013b) which provides real-time recognition of activities to disabled people. In addition to stream parallelism, Scribe (Lasecki et al., 2013a) also implemented a technique known as *warping time* which allows workers to listen to audio streams at reduced speeds to make it easier to carry out transcriptions or translations. Time warping was shown to improve precision and recall scores in real-time task. Other approaches in literature include a form of *Map Reduce* used in CrowdForge (Kittur et al., 2011) and the use of recursive workflows by Kulkarni et al. (2012). However, one additional approach which presents significant promise in carrying out paid microtasks within bounded time constraints is the use of *contests*.

3.2.3 Crowdsourcing Contests

The history of contests probably dates back to 1714 when the British Parliament ran one to determine the longitude at sea to within half a degree (Moldovanu and Sela, 2001). Galton in 1902 (Galton, 1902) then famously posed the problem of optimally dividing prizes in a competition, which was solved and has been proved theoretically by Moldovanu and Sela (2001) among others. Other well known competitions such as the Netflix Challenge (Bennett and Lanning, 2007) have been used to elicit a single best solution to a requester task. In the following we give an overview of empirical studies and theory in this space, which informed our model.

Researchers have investigated how contents unfold on existing platforms such as TaskCN (Liu et al., 2011a) and TopCoder (Yang et al., 2008). These studies have shown the effects of increased payoff as an indicator of contestant performance. The behaviours of contestants in TopCoder were analysed, for instance, by Archak (2010). In a related study Boudreau et al. (2011) considered over $9,000$ contests hosted on the platform in order to understand the effect of participant numbers on the performance of individuals. A second group of empirical work focused on bespoke experimental setups. For example, Rokicki et al. (2014) looked at the effect of varying monetary schemes and information policies in individual contests, while Rokicki et al. (2015) explored the same problem alongside team formation strategies in group-based contents. In this chapter we design and carry out contest experiments on our own platform as well. We study the effects of competition and exit patterns in order to run more effective paid microtasks projects that are time-sensitive.

Economists have extensively researched the theoretical foundations of contests. Recurring themes in this context are the optimal design of such contests (Archak and Sundararajan, 2009; Chawla et al., 2015), and the optimal allocation of prizes (Moldovanu and Sela, 2001). Others have look at payment mechanisms such as lottery contests (Rogers, 1998) and all-pay auctions (DiPalantino and Vojnovic, 2009). In situating our work in this space, our contest involved endogenous entry (Ghosh and McAfee, 2012) (as opposed to pay-to-join contests), and, it was a rank-order contest (Ghosh and Hummel, 2015; Lazear and Rosen, 1981) (as opposed to winner-takes-all contests), wherein we had access to cardinal information (Ghosh and Hummel, 2015) in the form of an absolute measure of the quality of each worker's submission. A survey of experimental research of contests is available from Dechenaux et al. (2014).

Finally, our work builds on literature which studies the *war of attrition.* In contests, each participant enters knowing their own skill and costs, but not that of the other contenders. Participants consider dropping out when they learn their opponents' strengths and discover that staying would be unprofitable. A theory of how this phenomenon operates in duopolies was presented by Fudenberg and Tirole (1986), while Krishna and Morgan (1997) presented its relation to an all-pay auction. On another landscape, Norrander (2006) discussed how strategic considerations such as assets, costs, and initial contest outcomes, could lead some candidates in political primaries to exit the race early. This also bears similarities to our scenarios in which contestants strategically decide whether they continue to take on more tasks or leave Wordsmith. Norrander (2006) introduced a duration model, which we also attempt to analyse in our work, which shows the length of candidacies and factors associated with candidate exits. Moldovanu et al. (2012) looked at contests with exits where contestants have the option to dropout or not to participate with the introduction of costless punishments. Since our objective was to maximise the total utility generated in real-time, we did not use punishments, which some early experiments we carried out revealed to increase the attrition rate.

3.2.4 Challenge

In our research, we focus on timely task completion, adopting existing methods in literature to obtain timely worker recruitment. One of the main challenges with ensuring real-time tasks are completed on time is maintaining and utilising a large workforce. This is coupled with the design of redundant workflows to ensure that no section of the task stream gets unaccounted for. In addition to this, crowd workers have varying task capabilities – most noticeably in real-time tasks – hence techniques like rapid refinement (Bernstein et al., 2011) make affordances for slower workers. Lasecki et al. (2014) also encountered this, recruiting a large workforce, but tasking only the fastest workers. What we seek is an approach to maximise the output of the fastest workers without the overhead of recruiting and compensating slower workers. Not every worker would be attuned to carrying out pressure-driven real-time tasks, and it is essential to match workers to tasks that play on their strengths. Getting slower workers to carry out real-time tasks would not only affect the quality of the task output, but could also hurt the worker (for example, they might get flagged or blocked by the requester). Therefore, we seek to address the challenge of real-time crowdsourcing by designing new or re-purposing existing techniques that would be attractive to fast workers and not create a cost burden for the requesters.

3.3 Motivation and Rewards

In this section, we briefly review some of the most relevant prior work pertaining to maximising the effectiveness of incentivised crowdsourcing. In particular, we focus on approaches that use game mechanics in human computation, and on methods that aim to optimize the performance of crowd workers, be that by offering bespoke incentives, or by assigning tasks to those workers who are likely to be able or willing to complete them accurately. As much of this background literature is inspired by, and explained using, theories of human motivation, we touch on fundamental work in that space as well.

3.3.1 Alternatives to Paid Microtasks

Alternative methods of crowdsourcing judgements have been reported in literature. Approaches that differ from the traditional methodology of eliciting judgements via an open call have been studied in a bid to improve paid microtasks, without jeopardising the quality of the work results.

Paid Microtasks vs Online Staffing

Online staffing platforms such as eLance[1] and oDesk[2] (recently merged into Upwork [3]) have served as a source of mid-term engagement of workers. Unlike traditional crowdsourcing platforms which are well suited to short bursts of microtasks, or open innovation platforms like InnoCentive [4] (where tasks can run for months), workers in a platform like oDesk are engaged for a few days (Nickerson, 2013).

An article by Ipeirotis (2012) details a comparative analysis between Mechanical Turk and oDesk. They highlight the suitability of oDesk to tasks that required longer term engagement, could be performed by fewer people, contained hard tasks that might be skipped by turkers, had to be completed within time (avoiding the long tail of MTurk) and required less anonymous workers.

Paid Microtasks vs Social Networks

Social networks such as Facebook and Twitter have been explored as sources of human agents for crowdsourcing tasks. Difallah et al. (2013) proposed an approach where workers are known and profiled in advanced (as opposed to the faceless crowd in traditional systems). Tasks are then pushed to these selected users as opposed to the traditional pull mechanisms.

Bozzon et al. (2013) considered the problem of choosing the right crowd by ranking the users of a social network based on their domain knowledge. The top expert users are then selected to solve the task at hand. Like Difallah et al. (2013), the approach requires profiling the users to find suitable candidates for the tasks. The Annotation-Validation (AV) Model: Rewarding Contribution Using Retrospective Agreement (Chamberlain, 2014b)

Paid Microtasks vs Games

Games have increasingly been used as a platform to engage crowd workers. Games With A Purpose (GWAPs) (von Ahn and Dabbish, 2008) were among the first to systematically apply game mechanics to create a fun environment for crowdsourcing tasks. The ESP game, (von Ahn and Dabbish, 2004) for example, motivated players to annotate images with descriptive tags, through a competitive framework in which they were pitted against other players to try to guess others' annotations as quickly as possible. Not

[1] https://www.elance.com

[2] https://www.odesk.com

[3] https://www.upwork.com

[4] http://www.innocentive.com/

only did this framework compel participation through direct competition, but it generated tags that were of high quality by directly rewarding consensus. Other similar image annotation gamified task environments included *Phetch* (Von Ahn et al., 2007) and *Peekaboom* (Von Ahn et al., 2006).

Thaler et al. (2012) presented a comparative analysis of user behaviour on MTurk and a GWAP. They compared the results of conceptual modelling and ontology mapping using a traditional crowdsourcing approach built on MTurk, and a game based approach using a custom designed game called OntoPronto. They reported significantly more results from the game platform at no cost per annotation; however, the average number of correct answers per participant in the game was significantly less than that in MTurk (8 vs 45).

Jurgens and Navigli (2014) reported a comparison of image annotation tasks. They compared the results of mapping senses from WordNet to images using a game called Puzzle Racer, and a traditional crowdsourcing system built on CrowdFlower. They were able to achieve comparable quality from both systems while reducing the cost by 63% by using the game based approach. However, the cost savings were at the expense of timely completion. Results from CrowdFlower were completed in hours, while results from Puzzle Racer trickled in over 2 weeks.

Eickhoff et al. (2012) carried out a study of crowdsourced judgements on relevance assessments and clustering. They presented evaluations using traditional HITs and gamified HITs on quality (compared against a gold standard), efficiency (time required to collect judgements) and incentives (financial vs fun). Their results show that with the gamified HIT, they were able to obtain quicker judgements at a higher quality by leveraging game flow and immersion as opposed to financial incentives in the traditional HITs.

Other relevant studies in literature include a comparison of crowd-Based, game-Based, and machine-based approaches by Harris and Srinivasan (2013).

3.3.2 Motivation and Incentives

The theory of motivation and an understanding of incentives is fundamental in understanding the *why* of worker behaviour in crowdsourcing (see Section 2.3.3).

While efforts at designing successful crowdsourcing projects have considered a variety of dimensions, including end-user interfaces, spam detection, and quality control, some of the most influential works in recent crowdsourcing literature have approached this problem by looking at crowd engagement. This is seen as an effective way to achieve better productivity and ensure the sustainability of crowdsourcing platforms over time. Research on crowd engagement covers various aspects, from studies of motivations of contributors to specific projects to applications of theoretical models from economics to the newer scenarios of online labour markets.

Motivation

The concept of intrinsic motivation emanated from the work of White (1959) in 1959. This was in contrast to the drive (drive-reduction, drive-induction) theory and instinct theory of that time. *Effectance motivation*, as it was called, could explain behaviours that did not require reinforcements and physiological drives, and encompassed learning, development, play, exploration and volitional behaviour. This was expanded to intrinsic and extrinsic motivation following the Self Determination Theory in 1985 by Deci and Ryan (1985b)

In the context of crowdsourcing Kaufmann et al. (2011) presented an extensive work on constructs of extrinsic and intrinsic motivations in crowd workers. They classified intrinsic motivation as enjoyment and community based, while extrinsic motivation as based on payoffs and social motivations. They observed that extrinsic motivation affects the length of time spent on the platform while intrinsic motivation (such as task autonomy) serves as the dominant factor for most workers.

A study of motivations of citizen scientists by Raddick et al. (2008) yielded 12 core motivation dimensions including fun, learning and discovery. Another study by Rogstadius et al. (2011) suggested that intrinsic motivation can increase the quality of workers' output by presenting tasks as helping others (e.g., helping a non-profit study malaria). All these works posited that contrary to the presented belief that money drives crowdsourcing or fun engages crowd workers, worker motivation transcends fun and money as aptly titled by Kaufmann et al. (2011).

Incentives

Yet for many classes of tasks, especially paid microtasks, the primary incentive for getting a crowd worker engaged is through cash payoffs. Optimising crowd payments have been studied in various forms including quota systems, performance based systems, studying target earners and using reservation wages (Horton and Chilton, 2010). However, financial incentives have been shown to improve speed of completion of tasks and not result quality by several studies including Mason and Watts (2010); Mao et al. (2013a); Yin et al. (2013). This has also led to studies on the anchoring effect where workers feel they should be paid more than they actually received (Mason and Watts, 2010; Yin et al., 2013) and the drop-off effect where workers stop working after hitting specific targets (Mason and Watts, 2010).

An extensive review of 74 experiments by Camerer and Hogarth (1999) present the effects of financial incentives in experiments, highlighting points where it helps, hurts or has no effects on mean performance. Their modal result showed that financial incentives had no effect on mean performance, however higher payments reduced the variance in results. In other instances such as cognitive tasks which were more responsive to better efforts, higher payments led to better judgements i.e., tasks which are easy require little capital, so paying extra won't help and vice versa.

Other forms of financial incentives have been adopted in crowdsourcing tasks. These include negative financial incentives where a portion of income is withheld for inaccurate results (Shaw et al., 2011; Harris, 2011).

3.3.3 Gamification

According to Zichermann and Cunningham (2011), gamification is 'the process of game thinking and game mechanics to engage users and solve problems'. Gamification leverages on and is different from games. The goal of gamification, which is essentially 'the use of game design elements in non-game contexts' (Deterding et al., 2011b), is to achieve a level of engagement seen in successful video games by transplanting some of the game elements (as opposed to play which has no formal rules) (McGonigal, 2011), design and mechanics to non-game tasks (Deterding et al., 2011b). Gamification often includes adding game-like rewards, and may also include competitive and social elements, such as leaderboards, explicit competitions, and group and individual performance feedback.

Gamification in Application

Many projects have already demonstrated substantial success in applying this idea to crowdsourcing settings; for example, the set of projects associated with Games with a Purpose (GWAPs) (von Ahn and Dabbish, 2008), have included the ESP game (von Ahn and Dabbish, 2004), a competitive image-tagging game that simultaneously created useful image labels for large image datasets. Other image labelling games that have followed include *Phetch* (Von Ahn et al., 2007) and *Peekaboom* (Von Ahn et al., 2006).

Perhaps the most salient example of a successful GWAP is *Duolingo*, a language-learning game that simultaneously helps players learn a new language, and to translate previously untranslated texts on the Web to other languages. Duolingo has become one of the top downloaded mobile apps of all time and still tops the educational apps charts on major mobile app stores [5]. Other highly visible gamified crowdsourcing projects are from *citizen science*, in which volunteers help complete large-scale scientific contributions. Both *FoldIt* [6] , a protein-folding game, and *EyeWire* [7] have seen massive sustained engagement and have contributed to new scientific discoveries more effectively through the application of gamification.

Gamification and Overjustification

Despite the successes seen with gamification, in some contexts, it has been seen to undermine intrinsic motivation by subjugating and trivialising contribution into simple

[5] Duolingo - a Visual History - https://www.duolingo.com/comment/3412629

[6] FoldIt - http://fold.it

[7] EyeWire - eyewire.org

game goals and points [8]. This effect has been called *overjustification* and has been demonstrated in a few studies, such as by Lepper et al. (1973), an experiment in which it was demonstrated that children that expected a reward performed more poorly than those who were not expecting any and were playing for purely intrinsic benefit.

Gamification and Extrinsic Rewards

Nonetheless, the studies of overjustification illustrate that the motivations for participating in various systems are both many and varied, and the effects of applying extrinsic rewards in various forms can help or hinder depending on the context. While the effects of overjustification have been reproduced, its prevalance seems to be highly dependent on the context; for example, a comprehensive survey by Deci et al. (1999) showed that in in a majority of cases, extrinsic rewards complemented, rather than undermined intrinsic motivations for participating. Similarly, another well-cited study showed that blood donations dropped when monetary rewards were introduced; and yet such overjustificaiton effects once again diminished when participants were allowed to donate their rewards to charity (Mellström and Johannesson, 2008).

3.3.4 Making Crowdsourcing Effective

While in the previous section we primarily looked at prior studies of motivation and incentives aspects in particular scenarios, we will now give an overview of methods which help make crowdsourcing projects more effective by optimizing key components of such projects. We posit that it is possible to align gamification incentives with gameful intrinsic motivations to yield maximal player engagement and quality player output.

Incentive Design Mechanisms

A number of descriptive frameworks have been proposed in the literature to capture the nuances of incentives engineering beyond simplistic *'fun or money'* considerations. Some of these include *MICE* (Money, Ideology, Coercion, Excitement) (Burkett, 2013), *RASCLS* (Reciprocation, Authority, Scarcity, Commitment, Liking, Social Proof) (Burkett, 2013), and *SAPS* (Zichermann and Cunningham, 2011). SAPS represents Status, Access, Power and Stuff, intended to represent a system of incentives from the most desired to the least desired, and the cheapest to the most expensive. We adopted this framework in our experiments.

Mechanisms for effective allocation of incentives have been studied in market and auction platforms, wireless and peer-to-peer networks and corporate organisations. In the context of crowdsourcing, a number of studies have been carried out, applying *game-theory* techniques to incentive design (Xie and Lui, 2014; Yang et al., 2012). These two pieces of work focus on financial incentives and a premise of inter-player strategy dependency. Not all crowdsourcing tasks can be modelled in this way; we adopt a probabilistic approach

[8] Criticisms of Gamification - `http://radar.oreilly.com/gamification-criticism`

based on prior player behaviours to predict appropriate incentives beyond the purely financial. Similar techniques are used for various purposes in crowdsourcing design, in particular to inform the assignment of tasks to workers or to predict task completion (Demartini et al., 2013; Sheng et al., 2008).

A large body of work has been dedicated to task and workflow design, as well as quality control (see, for instance, Michelucci (2013) for a recent compilation). We take their findings into account when implementing the basic interfaces published on CrowdFlower as well as the means to check quality and validate results.

3.3.5 Challenge

In our research, we seek to tackle the question of motivation and incentives, with particular interest in the case of worker drop off (Mao et al., 2013b) i.e., why do workers stop a task, and how can we motivate them to carry on a task after they attempted to quit. We call this incentive scheme *furtherance incentives*. Put succinctly, we are interested in designing furtherance incentive mechanisms that improve worker engagement and task uptake, while maintaining output quality and cost implications.

3.4 Synchronous Collaboration

Paid microtask crowdsourcing is essentially designed as an individualistic system. According to Gao et al. (2011b), the primary reason why crowdsourcing currently falls short for disaster relief is because: 'current applications do not provide a common mechanism specifically designed for collaboration and coordination between disparate relief organizations'. Even though crowdsourcing is meant to leverage on the wisdom of the crowd, individuals in the crowd are isolated with no platform for group formation or communication.

If interaction and collaboration yields positive benefits in traditional work and research contexts, the question from Kittur (2010) then is: '*would workers participating in a financial market really help each other without any financial incentives?*'. They carried out an experiment where a translation task was assigned to crowd workers. The workers were each paid $0.15, allowed to chat with each other, and permitted to see each persons translation in real-time. At the end of the task, 14 out of 16 bilingual speakers rated the crowdsourced translation higher than a published translation. The task also gave insight into which parts of the text were more difficult to translate – marked by the number of iterations by individual workers. Most importantly and surprisingly, at the end of the task, crowd workers set up a new translation task, which other workers joined in to translate for free. This experiment reflects a clear benefit of adopting collaboration. It is useful to reiterate that this experiment was carried out in a paid microtask setting where money serves as the primary incentive. It can be argued that the experiment task was creative in nature – and creativity flourishes with interaction and collaboration. Can collaboration really be integrated into paid microtask crowdsourcing as a way to harness the potentials of people working together?

3.4.1 Collaborative Crowdsourcing

Microtask crowdsourcing is usually modelled as an aggregation of individuals acting unilaterally. Workers act alone without interaction with other workers, while their individual outputs get assembled into a unit for the requester. This applies to paid scenarios (e.g., on CrowdFlower), as well as non-paid ones such as citizen science projects or games with a purpose (GWAPs). Nevertheless, other models have been introduced to build a community and facilitate crowd learning (Tinati et al., 2015); incentivize participation (e.g., via contests or social flow) (von Ahn and Dabbish, 2004; Rokicki et al., 2015); or support more complex types of tasks (Kittur et al., 2011).

Some GWAPs, for example, von Ahn's ESP game (von Ahn and Dabbish, 2008, 2004) and TagATune (Law and von Ahn, 2009) have a strong element of interaction among contributors. These games employ various strategies such as output agreement (ESP game) or input agreement (TagATune) between players to generate useful work results and drive engagement. Participants in these systems are primarily intrinsically motivated, either by their love for music, or by their desire to have fun - and not motivated (primarily) by money as with workers in paid microtask crowdsourcing (von Ahn and Dabbish, 2008, 2004).

In citizen science the interactive element has mostly a community building function. For example, Zooniverse projects such as Galaxy Zoo or PlanetHunters use discussion forums to allow contributors to ask questions, engage with other members of the community, or autonomously identify new lines of scientific inquiry to pursue based on previous observations in the data (Raddick et al., 2009; Tinati et al., 2015). Tasks are still carried out independently, though contributors can add comments and raise questions about a particular task, which are shared with the community. Answers cannot be revised based on these interactions.

Just as community spaces in citizen science led to serendipitous discoveries (Raddick et al., 2009), (Kittur et al., 2011; Kulkarni et al., 2012) demonstrate how involving the crowd in collaboratively designing the crowdsourcing workflow can have positive results on task quality and engagement in paid microtask settings. The same trend has been noted in Rokicki et al. (2015), where the authors present strategies for group formation in team-based crowdsourcing where members had to work together to perform image annotation. Finally, collaborative crowdsourcing has been promoted through an event that has received considerable attention in the mainstream media – the DARPA Red

Balloon Challenge. In the challenge teams had to find ten red weather balloons spatially distributed at undisclosed locations around the U.S., with over 4,000 people registering, extending to a group network size of over 350,000 participants. Tang et al. (2011) discusses the most successful strategies that were applied by challenge contestants, which, included, among other things, collaborative elements based on well-aligned incentives.

3.4.2 Factors Affecting Collaborative Participation

In most group settings, there is the potential for problems stemming out of lack of individual motivation. The free-rider effect (Kandel and Lazear, 1992) and social loafing (Huang and Fu, 2013) are among the issues which occur when the output of the group is considered as a whole without evaluating the contribution of individuals. On a positive note, however, peer dependency can lead to social facilitation and altruism (Huang and Fu, 2013). Social facilitation generally occurs when the contributions of individuals are evaluated, and compared with the contributions of other members of the group, while altruism stems from the desire to collaboratively affect other members of the group positively.

Social Pressure

Several works have looked at the effect of peer and social pressure in incentivising work output collaboratively online and in business enterprises. Directly relevant for our experiments were two broad sources of social incentives: empathy, guilt and shame (social pressure), and the desire to attain and re-experience social flow.

Kandel and Lazear (1992) presented their work on partnerships, and the incentives that can be generated through peer pressure wherein they highlight the role of empathy, guilt and shame, as incentive generators. They note that empathy generates incentives when one individual can positively affect the outcome of another individual's rewards. Peer pressure via empathy therefore becomes effective when all team members are either in shared circumstances - are at similar levels, and have potentially similar payoffs, or a better off worker can affect the income of another. Guilt and shame on the other hand are a function of transparency and observability of a worker's action by other members of the group, especially when the action has a group effect. More on the role of transparency was presented by Mohnen et al. (Mohnen et al., 2008) where they pointed out, just like Kandel and Lazear (Kandel and Lazear, 1992), that unobservability of contributions yields selfish agents (selfish people). A worker might feel internal guilt at not pulling their weight when their contribution is not visible to others, however, in a transparent setting, shame sets in, therefore putting pressure to perform more work. The effects of peer pressure on contributions to enterprise social media were studied by Brzozowski et al. (2009) where they observed that, the participation of a worker's manager is a key source of social pressure in initiating contribution while, comments fuel the pressure for sustained contributions. From their work, we observe that most of the social pressure effects that are seen in the offline world are present, and amplified in the online world.

Several articles also exist in theoretical economics literature, which model the cost of one person exerting pressure on another, which could then either lead to a reduction in the cost of the pressured person taking the action, or increase the cost of their not taking the action (positive and negative pressure) (Daido, 2004, 2006; Calvó-Armengol and Jackson, 2010).

Social Flow

Social flow stemmed out of an extension of Csikszentmihalyi's 'theory of optimal experience' (Csikszentmihalyi, 1991) where flow (or individual/solitary flow), was presented as an intrinsically rewarding, highly absorbing state, which is attainable when individuals freely choose an activity with: clear goals, immediate feedback, and a balance between challenge and skills. Despite the freedom and pleasure that comes from immersive individualistic activities, it has been observed that some of the most gratifying flow experiences occur in social experiences (Jackson and Csikszentmihalyi, 1999; Mockros and Csikszentmihalyi, 2014), leading to the concept of *social flow*.

Conditions
– The unit of performance is a group or team
– The collective competency of the group is sufficient to dispatch challenges
– Group members are uniformly highly competent
– Group members have task-relevant knowledge & skills about each other
– Emergent challenges are important & meaningful to the entire group
– Tasks prescribe interdependence, coordination & cooperation
– Tasks are conjunctive and require complementary participation
– Group members focus on each other as well as the task to receive feedback
– Task feedback is clear & immediate
– Task feedback is primarily cognitive and secondarily affective
– Social process feedback is primarily affective and secondarily cognitive
Indicators
– Shared intense absorption & engagement with the task
– High attention to group members or teammates
– Loss of sense of time
– Less awareness of self
– Surrender of self to the group
– Emotional communication during group work
– Emotional contagion within the group and observers external the group
– Joy, elation and enthusiasm felt and shared throughout group performance
– The experience builds meaning and a collective sense of purpose
– The group desires to the repeat the experience
– Rituals may be established to institutionalise social flow

TABLE 3.2: Conditions and indicators of social flow by Walker (2010)

Walker (2010) identified instances where co-active and interactive social flow is present such as: skiing down a mountain in a group and watching TV with buddies (co-active)

Individual solitary flow
– Doing work on my computer late at night. – Singing by myself in the car. – Composing choral music. – Painting with watercolors. – Gardening on a Sunday morning. – Cycling alone over rolling hills. – Running alone along the river as the sun rises. – Cooking by myself, home alone. – Writing a poem in the solitude of my familys cabin. – Reading a great book and relaxing in a hot bath.
Co-active social flow
– Running a marathon in a pack with others. – Skiing down a mountain in a group. – Playing golf with friends. – Hiking up a mountain with an outdoor club. – Listening to music with friends. – Watching TV with buddies. – Doing errands with friends. – Just sitting at the mall with friends watching people. – Cleaning while listening to NPR with my roommates. – Competing at a swimming meet.
Interactive social flow
– Playing soccer on a great team. – Joining a jam session at my neighbourhood jazz club. – Eating, drinking and talking with friends. – Exchanging funny stories and laughing with friends. – Having sex anytime with my lover. – Playing a game of pickup basketball. – Acting in a play on a night when everyone is on. – Having a heart-to-heart with a close friend. – Singing in a choir. – Ballroom dancing.

TABLE 3.3: Examples of individual and social flow by Walker (2010)

or playing soccer in a great team and ballroom dancing (interactive) (see Table 3.3). Walker also presented the conditions and indicators of *social flow*. Some conditions stated include: immediate and clear feedback from the task and group members, interdependence and cooperation, and the challenges are important to the whole group. Some indicators stated include: shared absorption and engagement, less awareness of self, and the desire to repeat the experience (see Table 3.2). In collaborative microtask crowdsourcing, the desire to attain and re-experience the social flow from solving a series of tasks together, serve as an incentive mechanism leading to improved task output.

3.4.3 Challenge

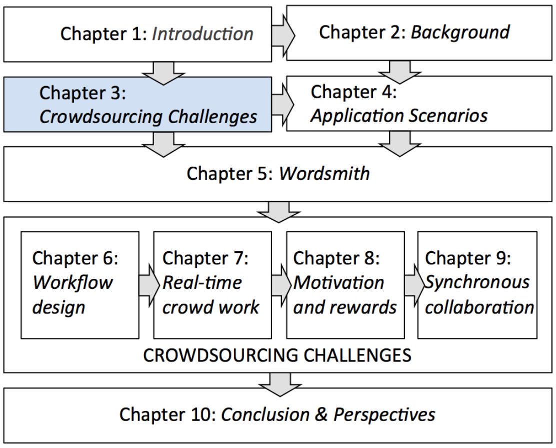
Our work was informed and inspired by the background literature just discussed. We developed a paid microtask environment, which recruits participants from platforms such as CrowdFlower, pairs them randomly as they log in to the system, and asks them to label images consensually.

The task design bears resemblance with multi-player GWAPs such as the ESP game, though the motivation of the participants and the aims of our experiments are different. When creating Wordsmith, our primary focus was not on coming up with a fundamentally novel game experience to collect image labels, but on building an experimental framework to test our research hypotheses regarding the interplay between monetary rewards, collaborative task design, and social pressure, and social flow. This is reflected in the experimental setup, which looks at task accuracy and output in three conditions: the traditional, single-worker one and two collaborative ones, one with and one without socially motivated incentives. Compared to previous work in paid microtask groupsourcing, we use a different collaboration model, building pairs of workers who complete tasks simultaneously, and study the effect of empathy-centric social pressure and social flow on crowd behaviour.

The challenge we seek to address here is to give a more comprehensive answer to the question posed by Kittur (2010). This requires further insight into the dynamics within groups, teams and collaborating workers – especially, to what level one worker can influence another to carry out a task. However, with all the benefits of group collaboration and coordination, negative issues such as the free rider problem and social loafing could still affect teamwork. The free rider problem (Kandel and Lazear, 1992) occurs when some team members do not play their part in achieving the overall goal, and still partake in the overall compensation – leading to an eventual decrease in output quality caused by the dissatisfaction of the contributing members. Similarly, social loafing (Latane et al., 1979; Karau and Williams, 1993) occurs when individuals exert less effort when working in a group than they would when performing an individual task.

These issues could be compounded in paid microtask environments where the anonymity of the crowd could further facilitate a non-chalant attitude to the task at hand. We believe curbing this, and answering Kittur (2010) question would necessitate studying socially motivated incentives. Social incentives such as peer pressure (Kandel and Lazear, 1992) have been observed to have effects in traditional organisations when applied horizontally between colleagues and vertically from bosses to subordinates. Similarly, social flow (an extension of Csikszentmihalyi (1991) theory of individual flow) describes how participating in group tasks could create a level of immersion that could not be attained alone. In this work, we seek to apply the concepts of social pressure and social flow to design collaborative crowdsourcing systems that improve worker engagement and task quality.

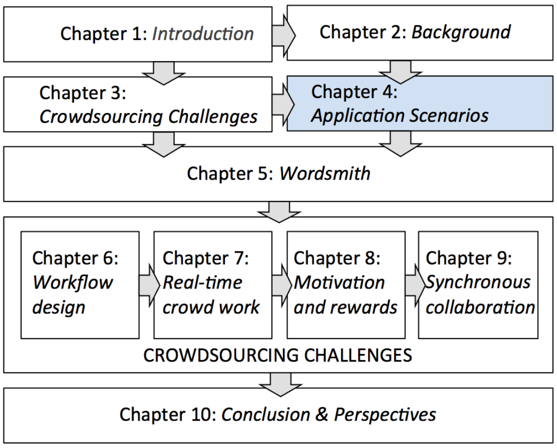
3.5 Summary


Chapter 1: Introduction
Chapter 2: Background
Chapter 3: Crowdsourcing Challenges
Chapter 4: Application Scenarios
Chapter 5: Wordsmith
Chapter 6: Workflow design
Chapter 7: Real-time crowd work
Chapter 8: Motivation and rewards
Chapter 9: Synchronous collaboration
CROWDSOURCING CHALLENGES
Chapter 10: Conclusion & Perspectives

In this chapter, we gave a detailed overview into the four crowdsourcing challenges that form the basis for this thesis: workflow design; real-time crowd work; motivation and incentives engineering; and collaboration in paid microtask. We highlighted the components making up each high level concept and discussed the state of the art in current research to identify gaps where our work builds up on. Finally, we outlined the specific parts of each challenge that this thesis seeks to address in the subsequent chapters.

Chapter 4

Crowdsourcing Application Scenarios


Chapter 1: *Introduction*
Chapter 2: *Background*
Chapter 3: *Crowdsourcing Challenges*
Chapter 4: *Application Scenarios*
Chapter 5: *Wordsmith*
Chapter 6: *Workflow design*
Chapter 7: *Real-time crowd work*
Chapter 8: *Motivation and rewards*
Chapter 9: *Synchronous collaboration*
CROWDSOURCING CHALLENGES
Chapter 10: *Conclusion & Perspectives*

This chapter extends our introduction to the crowdsourcing challenges by exploring two broad application areas which form the basis of all our experiments in future chapters: text annotation and image labelling. The chapter presents related work in the line of our selected scenarios. It also serves as a literature review of the state of the art in implementing customised platforms designed to address specific pain points encountered in crowdsourcing. Later on in Chapter 5, we introduce our own custom built platform – Wordsmith, drawing ideas from the literature presented here.

4.1 Overview

Crowdsourcing has come to find application in various industries and across different scenarios (Brabham, 2013). Since the concept was first introduced by Howe (2006), it seems to have found global adoption; virtually any industry, science discipline, or public sector agency could tell a story about how they reached out to the wisdom of the crowds to improve their services and react more flexibly to customer demand, run comprehensive data collection and analysis projects, or collect ideas and views for a better informed policy making (Dawson and Bynghall, 2012). In this chapter, we look at two popular areas of crowdsourcing adoption, i.e., (i) text annotation; and (ii) image labelling. In each area, we identify different application scenarios: for example, crowdsourced text annotations (specifically, Twitter annotation) is widely used for disaster relief and crises management, while crowdsourcing image labels serves as a precursor for training computer vision algorithms. In line with the research agenda for this thesis, we then discuss challenges faced in these two application areas:

1. Text annotation
   - Challenge: Workflow design
   - Challenge: Real-time crowd work
2. Image labelling
   - Challenge: Motivation and rewards
   - Challenge: Synchronous collaboration

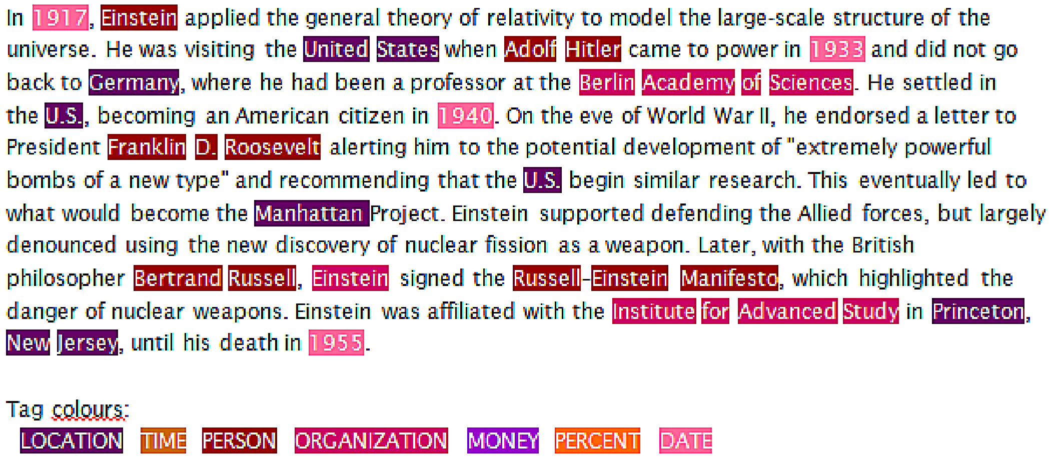
4.2 Text Annotation

Text annotation defined broadly would represent any form of explanatory markup attached to a part of the text to denote some referential meaning. For example, a piece of text in a sentence could either be annotated to represent a verb (part of speech tagging), or denote it as referring to a company (named entity recognition) such as is illustrated in Figure 4.1. The entire piece of text could also be annotated, for example, marking a piece of text as having a positive tone (sentiment analysis). One of the most notable text annotation projects was the Penn Treebank (Marcus et al., 1993), consisting of 4.5 million English words annotated for part-of-speech information. The initial project took at least 3 years (from 1989 to 1999). The underlying corpora in the Penn Treebank consists of well formed (grammatically correct) constructs of English text (e.g., sourced from the Wall Street Journal of 1987 - 1989). Furthermore, the project was carried out over a long period of time. However, the advent of the ‘Big Data’ age has made it essential to harvest, harness and annotate new forms of information – in large volumes and with speed. One familiar source of such information currently is Twitter.

The semantic analysis of microblog posts (or ’Making sense of microposts’, as a successful workshop series calls it)[1] has become one of the most active research topics in the Semantic Web area. With Twitter exceeding all predictions in terms of growth and influence,[2] analysing its vast amounts of user-generated data is essential for anyone aiming to gain a better understanding of how individuals, social groups, governments, and

[1] http://www.scc.lancs.ac.uk/microposts2014/

[2] 200 billion tweets per day, referenced by more than 1 million third-party websites, yielding over 30 billion impressions, according to their latest SEC filing. See Twitter Inc, form S-1 at http://www.sec.gov, accessed 2014 – 02 – 17.



In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell–Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:

Figure 4.1: Named Entity Recognition (http://www.europeana-newspapers.eu)

businesses communicate and interact online. However, it is also a challenging task, primarily due to the nature of the content (limited number of characters per post, extreme variation in writing styles, out-of-vocabulary words etc.), and the size and dynamicity of the datasets; all these aspects make the application of off-the-shelf Information Extraction (IE) tools, even when they offer support for semantic technologies or Linked Data, hardly feasible (Ritter et al., 2011; Finin et al., 2010; Derczynski et al., 2015).

Previous work in IE for microblogs has adapted traditional Natural Language Processing (NLP) techniques to reflect the specifics of micropost content. This refers mainly to the removal of stop words, retweets, hashtags symbols, ellipses, links, 'user' mentions, as well as out-of-vocabulary words (i.e., 'b4' or 'shuld') (Han and Baldwin, 2011). Other common approaches include text tokenization and optional parts-of-speech (POS) tagging, which use keyword selection to compute the 'link probability' (to Wikipedia article titles) of the tokenized text in order to identify potential entities (Sachidanandan et al., 2013). Similar methods resort to Wikipedia to match tokenized texts (Genc et al., 2013), as well as POS tagging to train and identify nouns to be further analysed (Muñoz-García et al., 2013). Alternative sources of keyword matching involve Freebase (Laniado and Mika, 2010), DBpedia (Jadidinejad, 2013; Mendes et al., 2013; Muñoz-García et al., 2013), and WordNet (Sachidanandan et al., 2013). The CMU POS Tagger has been developed to handle Twitter-specific vocabulary such as abbreviations (e.g., 'ikr', 'smh'), emoticons (e.g., ':o', ':/'), hashtags, and mentions (Das et al., 2013). Oliviera et al. (de Oliveira et al., 2013) used five filters (Term, Context, Affix, Dictionary, Capitalization) to decide upon potential entities over continuous Twitter streams. The approaches to building these automatic annotation tools have generally fallen in three categories (Liu et al., 2011b): (a) rule based; (b) adopting machine learning and (c) a hybrid of rules and machine learning. An analysis of the state of the art in named entity recognition and linking on using 8 tools was also presented by Derczynski et al. (2015). They also highlight the approaches used by these tools as: (a) using gazetteers (with finite state machines, or rules, or similarity metrics); and (b) using CRF and Machine learning. Building advanced modern tools using machine learning requires a training set of annotated data which has been labelled by humans.

However, given the volume of accessible tweet data, it is infeasible (on a time and cost scale) to have annotations carried out by a few experts. This underscores the case for leveraging microtask marketplaces and non-expert crowd workers as a scalable source of manpower to label tweets. One of the earlier works on crowdsourcing annotations in tweets was carried out by Finin et al. (2010). They used CrowdFlower and Amazon's Mechanical Turk as 'recruitment agencies' to source crowd-workers who were required to annotate occurrences of persons (PER), organisations (ORG) and locations (LOC) in a corpus of $4,400$ tweets.

One of the earlier works focusing on utilising the crowd for annotation tasks was by Snow et al. (2008) where they used a pre-computed gold standard to improve annotator quality. Several other approaches has been presented to improve the quality of task output by crowd workers. These include using detailed annotation guidelines; engaging multiple annotators (Lawson et al., 2010) and relying on results with high inter-annotator agreements. A set of guidelines for corpus annotation, distilled from existing literature was presented by Bontcheva et al. (2014a). Of note are the sections on *in-task quality*, *contributor evaluation* and *aggregation* where various approaches such as the use of gold standards, majority voting, active learning and average reliability are mapped to their adoption in literature. The role of uncertainty arising from worker annotation was addressed by Plank et al. (2014) by looking at inter-annotator agreement loss. Also of importance in crowdsourced annotation is the role of worker diversity (Trushkowsky et al., 2013) which improves recall by unearthing patterns which could not be seen by a homogeneous set of limited experts. Further factors also affect worker quality beyond the presence of a diverse crowd. Some extrinsic factors affecting annotation quality were presented by Cohn and Specia (2013).

In the next sections, we discuss application scenarios for crowdsourcing annotations on Twitter datasets.

Scenario: Crowdsourcing named entities

Several approaches have been applied to build tools for entity extraction, using rules, machine learning, or both (Liu et al., 2011b). An analysis of the state of the art in deploying software tools for named entity recognition and linking on microposts is available in Derczynski et al. (2015). The authors also discuss a number of factors that affect precision and recall in current technology - current limitations tend to be attributed to the manner of text e.g., vocabulary words, typographic errors, abbreviations and inconsistent capitalisation (Feyisetan et al., 2014; Ritter et al., 2011).

Crowdsourcing has been previously used to annotate named entities in micropost data in a study by Finin et al. (2010) which was introduced briefly in an earlier paragraph. In this study, Finin et al. used CrowdFlower and Amazon's Mechanical Turk as platforms. Crowd workers were asked to identify person (PER), location (LOC) and organisation (ORG) entities. Each task unit consisted of 5 tweets with one gold standard question, with 95% of the tweets annotated at least twice. The corpus consisted of 4, 400 tweets and 400 gold questions. Gold questions (gold data, gold standard) are questions with answers known to the task requester. This is used to evaluate worker performance and weed out spammers. A review of the results of Finin et al. (2010) was carried out and reported in Fromreide et al. (2014). They observed annotations that showed lack of understanding of context e.g., *china* tagged as LOC when it referred to *porcelain*. They also highlighted the issue of entity drift wherein entities are prevalent in a dataset due to temporal popularity in social media. This adds to the difficulty of named entity recognition (Derczynski et al., 2015) and highlights the challenge of solving this task using a simple crowdsourcing workflow.

A similar approach has been used to carry out NER tasks on other types of data. Lawson et al. (2010) annotated 20, 000 emails using Mechanical Turk. Their approach incorporated a bonus system which allowed the payment of a bonus in addition to the base amount contingent on worker performance. The workers were also required to annotate person (PER), location (LOC), and organisation (ORG) entities. By incorporating a bonus system based on entities found and inter-annotator agreement, they were able to improve their result quality considerably. The results were used to build statistical models for automatic NER algorithms. An application in the medical domain is discussed in Yetisgen-Yildiz et al. (2010). The crowd workers were required to identify and annotate medical conditions, medications, and laboratory tests in a corpus of 35, 385 files. They used a custom interface (just as we do in our experiments) and incorporated a bonus system for entities found. Voyer et al. (2010) presented a hybrid approach where expert annotators identified the presence of entities while crowd workers assigned entity types to the labels. This approach by Lawson et al. (2010) also used a simplified workflow while relying on the use of bonuses to improve annotation results.

Demartini et al. (2012) proposed a hybrid crowd-machine workflow to identify entities from text and connect them to the Linked Open Data cloud, including a probabilistic component that decides which text to be sent to the crowd for further examination. Using hybrid systems to offer crowd based query processing has also been studied by Trushkowsky et al. (2013). Their work leveraged on the crowd to improve recall scores in open-ended questions and how a mixed crowd can help converge on an accurate answer. Other examples of similar systems are Braunschweig et al. (2013) and Sabou et al. (2014). Sabou et al. (2014) also discussed some guidelines for crowdsourced corpus annotation (including number of workers per task, reward system, task quality approach, etc.,), elicited from a comparative study. A similar set of recommendations based on task character, human participation and motivation, and annotation quality was presented by Wang et al. (2013).

Compared to the works cited earlier, we performed quantitative analysis based on controlled experiments designed specifically for the purpose of exploring performance as a function of content and crowdsourcing features. The primary aim of our research (i.e., those that involved text labelling scenarios, and specifically, the workflow design challenge) was not to implement a new NER framework, but rather to understand how to design better hybrid data processing workflows, with NER as a prominent scenario in which crowdsourcing and human computation could achieve significant impact. In this context, our purpose built platform (introduced next in Chapter 5) is seen as a means to outsource different types of data-centric tasks to a crowd and study their behaviour, including purpose-built features for quality assurance, spam detection, and personalized interfaces and incentives.

Scenario: Sentiment analysis and opinion mining

Sentiment analysis refers to techniques employed to extract subjective information representing the emotional state (affective or intended) of the author of a piece of text. The sentiment might be represented at a high level as positive or negative, happy or sad; or in a more nuanced fined grained form as angry, depressed or excited. Understanding the sentiment in consumer comments has become important in the fields of brand management, marketing and public relations. Targeted sentiment analysis (Ghiassi et al., 2013) reveals opinions held not on the sentence level as a whole, but on a specific entity in the sentence. For example in the following sentence, the sentiment is negative, and directed at the entity 'Apple iPhone 6':

– *My **Apple iPhone 6** is all the bad things of all the different phone brands in one.*

Two of the earliest works on carrying out sentiment analysis and opinion mining on Twitter data were carried out by Go et al. (2009) and Pak and Paroubek (2010). Both approaches recognised the non availability of training data and therefore used emoticons as 'noisy labels' to pre-judge tweet sentiment i.e., tweets with ':-(', ':(', '=(', ';(' were assigned negative sentiments, while tweets with ':-)', ':)', '=)', ':D' were assigned positive tweets. These serve as a guesstimate on the sentiment of the tweet in the absence of hand labelled data. Other approaches such as by Kouloumpis et al. (2011), used word features to detect sentiment. However, with the growing utility of microtask market places, recent efforts such as SemEval-2013 Task 2 (Sentiment Analysis in Twitter) by Nakov et al. (2013) carried out the labelling of tweets via crowdsourcing. The crowdsourced corpus was then used to train and evaluate automatic sentiment analysis by 44 teams. The barrier to entry for these teams was therefore lowered by the availability of the annotated corpus.

Scenario: Breaking news and crises management

After the news of the death of Osama bin Laden leaked on Twitter (Hu et al., 2012) [3]

[3] http://www.huffingtonpost.com/2011/05/02/osama-bin-laden-death-twitter-leak_n_856121.html

when Keith Urbahn, former Defence Secretary Donald Rumsfeld's chief of staff posted a tweet, it became obvious that Twitter could serve as a veritable source of real-time breaking news. Another compilation of purported breaking news which first appeared on Twitter [4] includes high profile stories such as the death of Michael Jackson and the 2009 New York plane crash. This has led to numerous studies on Twitter as a source of breaking news (Phuvipadawat and Murata, 2010; Kwak et al., 2010; Vis, 2013; Hu et al., 2012; Petrovic et al., 2013) and how it might be harnessed as a journalistic reporting tool. However, as opposed to newswire outlets, posted tweets are anecdotal and at the very best, unverified pieces of information which might be downright rumours. For example, during the 2011 England riots (Vis, 2013), there were tweets of the London Eye being on fire. This turned out to be false. Studying these phenomena – that is, true and false news on Twitter, led to the creation of academic projects such as the Pheme Project [5]. One of the approaches employed in the project was crowdsourcing the annotation of rumourous conversations as reported by Zubiaga et al. (2015) which serves as a first step to detecting false information automatically. Breaking news on Twitter is usually accompanied by various hashtags, which gradually coalesce to a set which can be used to monitor events as they unfold. This has led Twitter once to be referred to as a medium for 'crowdsourcing the news' (Vis, 2013).

A specific class of breaking news, which finds particular prominence on Twitter pertains to natural disasters and national crises. The 2010 Haiti Earthquake was described by Heinzelman and Waters (2010); Forrest (2010) as the first disaster in which open-source and online platforms were heavily utilised. Four of such platforms included CrisisCamp Haiti, OpenStreetMap, Ushahidi (a specialised crowdsourcing platform), and GeoCommons which were reported by Zook et al. (2010). Of primary concern in the wake of a disaster is who needs help, and where, which informs the 'how' of channelling the help. Given the lack of geospatial information within and around Haiti at the time of the quake, one of the ways information on how to help spread was via Twitter. The #Haiti hashtag quickly spread on Twitter as a way to help gather information which was also further routed on SMS crowdsourcing platforms such as Ushahidi [6]. Ushahidi volunteers manually monitored the hashtag and were able to geolocate help requests on the ground. Since then, the Ushahidi platform in conjunction with tweet monitoring, has been deployed during other times of crises such as the 2010 Chile Earthquake and the 2010 Russian wildfires. Several researchers have looked into the role of Twitter as a vital tool for mining data during disasters with crowdsourcing as the principal tool of harnessing relevant information (Gao et al., 2011a; Goodchild and Glennon, 2010; Gao et al., 2011b; Heinzelman and Waters, 2010; Ortmann et al., 2011; Starbird, 2011). As with breaking news (of which disasters are a subset), the crowdsourced data can be used

[4] http://www.techradar.com/news/internet/10-news-stories-that-broke-on-twitter-first-719532

[5] http://www.pheme.eu/

[6] https://www.ushahidi.com

to train automatic tools for extracting information in real-time during crises (Kumar et al., 2011).

A characteristic of breaking news is the flurry of fragmented information that streams in as the event unfolds. The use of hashtags make it slightly easier to aggregate such information on Twitter, however, people tend to use different hashtags during the early and most critical moments of the disaster (e.g. #Haiti, #SaveHaiti, #PrayForHaiti). Apart from that, it is also essential to separate the commiseratory tweets, from those offering information on the ground which could assist in the relief efforts. Just as Ushahidi volunteers manually monitored hashtags, scaling out via crowdsourcing would require constant monitoring and real-time 'harvesting' of relevant information.

4.2.1 Challenge: Workflow Design

From the iPhone sentiment example above, we observe multiple steps required to obtain the final targeted sentiment: first, named entity recognition (NER), then the sentiment analysis task. How could we design this as a crowdsourcing task to leverage on the wisdom of the crowd? Given the initial task (named entity recognition), it becomes a non-trivial task when a group of non-experts attempt to annotate the text entity – is it *Apple iPhone 6* as a product, or *Apple* as a company and *iPhone 6* as the product, or simply *iPhone* as the product. A simpler example perhaps would be the sentence:

– *President Barack Obama is on his way to London, fun times ahead!*

This can easily be annotated by state of the art NER tools – recognising 'Barack Obama' as a person, and 'London' as a location with high precision and recall. It would thus be a waste assigning this sentence to be annotated by the crowd. A more tricky example noted by Derczynski et al. (2015) is:

– *Branching out from Lincoln park after dark ... Hello 'Russian Navy', its like the same thing but with glitter!*

Despite the capitalisation, 'Lincoln Park' does not refer to a location, however the compound phrase 'Lincoln park after dark' refers to a nail varnish colour. These 3 examples shed light on a possible workflow for crowdsourcing named entities on microposts. Certain tweets are essentially trivial and could be annotated with high accuracy using an automatic tool. Other tweets yield high consensus and can be outsourced to the crowd. While a final set generate disagreement and might need to be settled by a team of experts. Designing a workflow that plays to the strength of the crowd becomes essential in: minimising costs by annotating appropriate tweets using automatic tools, and maximising accuracy by utilising experts where necessary.

4.2.2 Challenge: Real-time Crowd Work

Microtask crowdsourcing has often been praised for its ability to produce results quickly and accurately. Yet, an increasing number of applications come in with much harder time constraints, which push the boundaries of the traditional microtask model to deliver in seconds or less (Bernstein et al., 2011). Examples of such real-time crowdsourcing applications include machine learning for image recognition (von Ahn and Dabbish, 2004) and text-to-speech conversion (Lasecki et al., 2013a); accessibility design (Bigham et al., 2010; Lasecki et al., 2012a, 2013a); and disaster management (Gao et al., 2011a).

Timely worker recruitment and task completion are, alongside better crowd engagement, key to mastering time-critical crowdsourcing scenarios (Lasecki et al., 2014). Several approaches have been proposed in the literature to address them (Lasecki et al., 2014, 2013a, 2012a; Bernstein et al., 2011; Bigham et al., 2010). However, they tend to use much larger crowds than necessary (Lasecki et al., 2014) or recruit in advance (Bigham et al., 2010) – in both cases the costs add up quickly for high volume, high throughput problem spaces.

The challenge then would be two-fold: (i) to recruit sufficient workers just-in-time at the outset of a breaking news report and (ii) employ a crowdsourcing design pattern that would facilitate real-time annotation of the streaming datasets. The difficulty of recruiting a large crowd on time is equally compounded by the cost of paying them per unit time as new pieces of work come in (which might not contain useful information – although, automatic filtering mechanisms can cut down on noisy inputs). The main challenge therefore involves undertaking real-time annotation, at reasonable accuracy levels, low cost, and within hard time constraints. Our aim was to come up with a proposal that is faster than existing approaches, while keeping the costs manageable. We believe these two aspects are vital in order for microtask crowdsourcing to establish itself as a data processing component that can be applied to large data sets to enhance automatic algorithms in real-time.

4.3 Image Labelling

Image labelling is an important pipeline component in artificial intelligence and computer vision research (Sorokin and Forsyth, 2008). Earlier image datasets were built in-house either by teams of experts manually assigning labels to images or through automatic techniques. Some of these include: the Berkeley database (Martin et al., 2001) containing natural images (a targeted total of 1,000 images), which were segmented by a group of individuals based on the subjects in the image; the Caltech dataset of 101 image categories (Fei-Fei et al., 2006; Griffin et al., 2007) which was trained on a few images labelled by 2 subjects; a dataset of 44,773 faces with an initial training set of 1,000 faces manually labelled by Berg et al. (2005); and a corpus of 13,233 target faces in unconstrained environments by Huang et al. (2007). The study by Huang et al. (2007) further surveys 30 datasets of face images with corpora sizes ranging from 185 (M2VTS Multimodal Face Database) to 99,000 (CAS-PEAL Face Database) images.

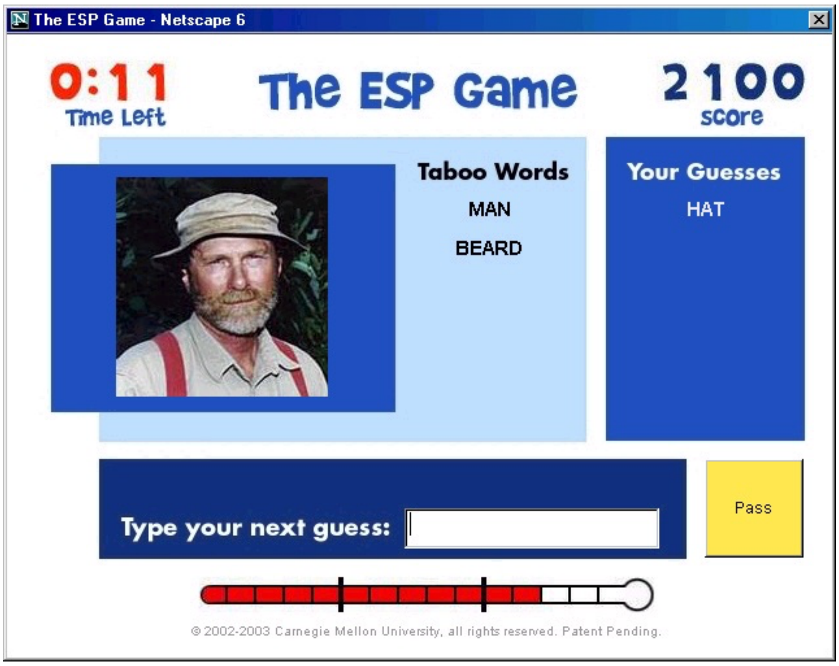

The ESP Game - Netscape 6
0:11
Time Left
The ESP Game
2100
score
Taboo Words
MAN
BEARD
Your Guesses
HAT
Type your next guess:
Pass
© 2002-2003 Carnegie Mellon University, all rights reserved. Patent Pending.

Figure 4.2: The ESP Game (von Ahn and Dabbish, 2004)

However, the practice of manually labelling images in a lab gradually gave way to other scalable approaches. One of such is LabelMe by Russell et al. (2008). LabelMe was an early web platform for annotating images to be used in object detection and recognition research. This helped to scale the labelling process beyond a group of localised annotators. Others were image annotations designed as games, such as the ESP game by von Ahn and Dabbish (2004) (shown in Figure 4.2) and Peekaboom (Von Ahn et al., 2006). The ImageNet project [7] (Deng et al., 2009) is an ongoing research project to organise images according to the WordNet hierachy [8]. WordNet is a lexical database of English words with synonyms grouped into synsets interlinked by semantic relations. ImageNet contains over 500 images for each noun based WordNet node, with a total of 14,197,122 images (with a target of 50 million) and 21841 synsets indexed at the time of writing. ImageNet sources its images automatically via search queries to search engines. However, the verification and annotation of the images into synsets is carried out by crowd workers on Amazon Mechanical Turk. Together with the work by Sorokin and Forsyth (2008), this set the scene for large scale, scalable annotation of image data by leveraging on crowdsourcing.

[7] http://image-net.org/

[8] http://wordnet.princeton.edu/

Building large annotated image banks like ImageNet come with challenges – even moreso when building an imageset for a targeted domain. ImageNet leverages on the maturity of existing image search engines, therefore creating a tag bank with low initial cost overhead. However, in order to build domain specific image annotations, or, for segmentation of image elements at scale, the search engine query would not be an option. In such cases, the images would have been generated as a result of an earlier domain phenomenon – e.g., medical images or deep space imagery. Annotating these niche images would require a large and *motivated* crowd as the first step. Afterwards, the annotations need to be verified to yield accurate labels. In the next sections, we discuss two application scenarios of image labelling. We also highlight challenges faced in using the crowd to carry out the annotations in each of the scenarios.

Scenario: Medical imaging

One of the domains where crowdsourcing finds niche application in is the sub-domain of medical imaging. Images are produced in clinical settings in vast numbers, providing critical information for diagnosis, treatment planning and other tasks (de Herrera et al., 2014). Radiology, endoscopy, magnetic resonance and radiography are but a few sources of images which require interpretation. A study by de Herrera et al. (2014) demonstrates the suitability of crowdsourcing for medical image classification. By using workers from CrowdFlower, they demonstrated that crowdsourcing could be used to improve the quality of an automatic classification task by increasing the amount of the training set. Crowdsourcing was used to create and correct the training set with strict quality control parameters at a 'very limited cost'. Other research works such as Mavandadi et al. (2012) and Luengo-Oroz et al. (2012) describe specialised crowdsourcing systems (designed as games) for identifying the presence of malaria in infected red blood cells.

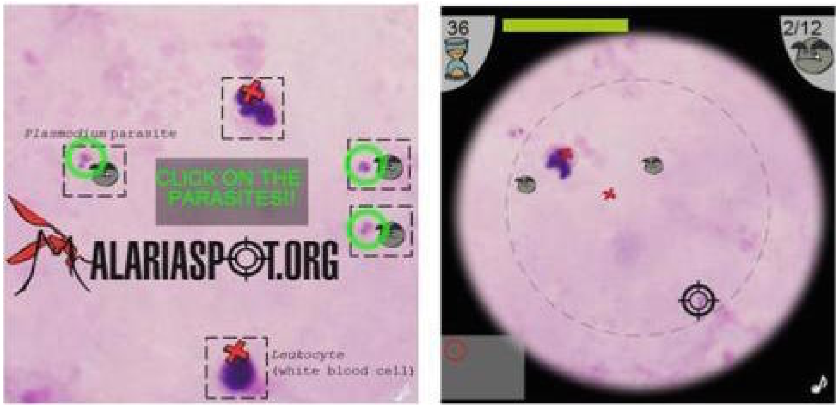


Figure 4.3: MalariaSpot (Mavandadi et al., 2012)

In these games (particularly MalariaSpot shown in Figure 4.3), 'untrained' crowd workers or casual game players are able to achieve annotation accuracies over 99%. This yields significant cost and time savings translating to actual saved lives. The challenge then is: how to get enough people to annotate malaria test images (or any other clinical image).

Scenario: Computer vision
In the book – '*Computer Vision: A Modern Approach*', Forsyth and Ponce (2003) presented crowdsourcing as a means of gathering data collections cheaply. The datasets created manually by researchers in the previous section, were all created to improve the computer vision capabilities of certain algorithms. According to Wah (2006), crowdsourcing finds its application in computer vision in the following respects: large scale data collection, image annotation, video annotation, investigating deficiencies and performing classification tasks that a difficult for computers.

4.3.1 Challenge: Motivation and Rewards

For domain specific image annotations, one of the primary challenges is attracting enough participants. For example, under the Zooniverse umbrella of citizen science projects (Raddick et al., 2008), ranging from astronomy, ecology, cell biology, humanities, and climate science, some projects have failed to reach critical mass, while others have simply not taken off as quickly as the more popular Galaxy Zoo projects. These domain specific labelling projects, just like medical imaging annotation projects, need to be designed in ways that motivate participation.

4.3.2 Challenge: Synchronous Collaboration

Crowdsourcing for computer vision serves as a first step for building better AI algorithms. As such the training data has to be as accurate as possible. This has led to the design of various workflows such as 'find-fix-verify' (Bernstein et al., 2010) which use multiple workers in sequence to ensure the quality of the task output. This is not unusual in actual work and academic circles with collaboration and interaction being a mainstay of social life. This has informed the idea of adopting synchronous collaboration as a way to improve the quality of crowdsourcing tasks, and serve as an incentive mechanism in its own right. However, given the individualistic nature of traditional paid microtask crowdsourcing, designing systems where workers become reliant on one another (especially for eventual financial payouts) presents unique challenges – including the potential to be completely boycotted by the workers.

4.4 Background and Related Work

In this section, we present related work in the lines of our two application scenarios i.e., text annotation and image labelling. We discuss specific crowdsourcing platforms from related literature, implemented within the context of the scenarios of interest. In Chapter 5 we introduce a our crowdsourcing platform (Wordsmith) which sources workers from existing microtask marketplaces. We designed Wordsmith, gleaning insights from the systems discussed in the following paragraphs and centering it around the four challenges we seek to address.

4.4.1 Text Annotation

Phrase Detectives - Poesio et al. (2015)
Phrase Detectives is a single player game for anaphora annotation and resolution. Anaphora is the linguistic mechanism of identifying an entity already used in a text. In Phrase Detectives (Chamberlain et al., 2008), players identify text as either referring to an earlier mentioned entity or a new entity. This is the annotation section of the game and it is called *name-the-culprit*. The validation section of the game, called *detectives conference*, presents a player with an annotation from another player. The player then either agrees with the submitted choice or switches to annotation mode to enter a new answer. Phrase detectives consists of a training phase with gold standard questions shown to new players; it uses a point system to offer feedback on correct answers; and it awards bonuses for agreeing with the gold standard on subsequent questions. It also advances players across levels with comparative scores displayed on a global leaderboard. We have also incorporated all these ideas into Wordsmith. Unlike Wordsmith, Phrase Detectives does not primarily source players from crowdsourcing marketplaces such as Amazon's Mechanical Turk or CrowdFlower. However, it has a Facebook version that sources non-anonymous players from the social network, leading to better quality control.

PlayCoref - Hladká et al. (2011)
PlayCoref is a single player and two-player game similar to Phrase Detectives. However, unlike Phrase Detectives which is played in single player mode and focuses on anaphora detection, PlayCoref focuses on detection of coreference chains. Hladká et al. (2011) gives 8 further differences between the two systems. In PlayCoref, players read a text document for 5 minutes and then connect all co-referencing words (as opposed to full phrases) in as many sentences as possible. During the game session, a player can see the number of words their opponent has linked into the coreferential pairs. The player also has access to the number of sentences with at least one coreferential pair marked by the opponent.

Dr. Detective - Dumitrache et al. (2013)
Dr Detective is a single player game that engages players into solving annotation tasks on medical case reports. Unlike most crowdsourcing platforms and GWAPs, Dr Detective players are medical experts (as opposed to untrained annotators), and, it is tailored to locate disagreements (as opposed to annotator agreement). Dr Detective allows users to carry out different kinds of annotation tasks on medical case reports including: term extraction (identifying relevant terms in a text), term categorisation (classifying a term

into an appropriate category), relation extraction (identifying whether or not a relation exists) and relation categorisation (classifying a relation into an appropriate category). It also features game mechanics such as point scores, levels and a leaderboard.

Sentiment Quiz - Rafelsberger and Scharl (2009)
Sentiment Quiz was one of the early games targeted at sentiment detection. It sources game players (and their social circle) from Facebook, who are required to evaluate whether sentences and dictionary terms express positive or negative sentiments. Players select a sentiment on a 5-point scale associated with a given word. The dataset used consisted of messages surrounding the US Presidential Election 2008 where people and media outlets expressed different views on contesting candidates. Sentiment Quiz's scoring is based on annotation agreement with disagreement leading to penalty points. It leverages on the Facebook platform not only to attract users, but also to promote scores, levels and leaderboard visibility.

PackPlay - Green et al. (2010)
PackPlay is a collaborative game for annotating semantically rich corpora. It consists of two games variants: *Entity Discovery* and *Name That Entity*. Players are paired with an anonymous partner (or a bot that mimics a previous player) when they begin the game. In order to maintain output quality, every player completes part of a 60 sentence pre-test drawn from a gold standard of know answers within the Tjong et al. (2003) dataset. The *Entity Discovery* game is a named entity recognition task where the paired partners are to anonymously and correctly identify as many entities (person, organisation and location) as possible. Scores are given for overlaps. The *Name That Entity* game is the verification stage of the game where players are always paired with a bot that shows selected entities from previous *Entity Discovery* game runs. The player is then required to select the matching entity type.

PhraTris - Attardi et al. (2010)
PhraTris is a game for annotating sentences with syntactic dependencies. The game bears semblance with the brick assembling game *Tetris* - hence its name. Rather than piecing bricks together, a player is required to rearrange blocks of sentences in a logical manner. PhraTris is a single player game without a strong crowd component. However, Its core can be re-purposed to build a collaborative or crowd powered game using features from other surveyed platforms.

4.4.2 Image Labelling

ESP Game - von Ahn and Dabbish (2004)
This is perhaps the most popular annotation platform which popularised the term Games with a purpose. The ESP game is an interactive two-player game in which two randomly paired are shown an image. The objective is to guess what labels the partner uses to

describe the image thereby advancing them to the next image (hence the name extrasensory perception or ESP). As the players advance to new images, they build up scores. From its launch in 2003, ESP has amassed over 200,000 players, annotating over 50 million labels. However according to Deng et al. (2009), in a bid to quickly move through images, players tend to annotate images with high level constructs such as *animal* or *dog* rather than *husky* or *greyhound*. ESP also does not afford for segmentation or identifying the position of objects in images which was addressed in further work such as Von Ahn et al. (2006)

Peekaboom - Von Ahn et al. (2006)
Peekaboom is an image annotation game designed by the authors of the ESP game to address some of its shortcomings especially that of object position identification. Peekaboom is also an interactive two-player game, however, unlike the ESP game, players have asymmetric roles. An image is presented to one player (called Boom), while the other player (called Peek) is to guess the image without initially seeing it. The *Boom player* hints the *Peek player* by clicking on segments of the image which are then revealed. Once the Peek player guesses correctly, the roles are switched and vice versa continue for four minutes. Peekaboom improves on the ESP game by providing not only image identification, but also *partial* image segmentation and information as to the position of the object in the image.

LabelMe - Russell et al. (2008)
LabelMe is a database an annotation tool used to carry out detailed annotation and segmentation of images. The images consist of the MIT CSAIL Database of objects and scenes [9], and other images taken by the authors, leading to over 14,000 images. LabelMe allows annotators to identify not only high level captions, but also objects embedded in a scene. Therefore, unlike the ESP game, and improving on Peekaboom, LabelMe offers position information, as well as segmentation in form of object shapes. The annotation incentive is access to the database, i.e., one has to annotate at least 10 images to have access to the entire dataset. LabelMe features a strong element of choice without being overly pedantic on the choice of labels used. However, this leads to a bottle neck as label verification is not done by the crowd, but by the authors. Unlike Wordsmith, LabelMe does not source its annotators from crowdsourcing marketplaces such as CrowdFlower or Mechanical Turk.

Phetch - Von Ahn et al. (2007)
Phetch is a multi-player interactive game for annotating images with accurate long form captions. Unlike the ESP game and Peekaboom, rather than labels such as *man* or *flute*, players come up with descriptive captions such as *an abstract line drawing of a man with a violin and a woman with a flute*. The game is played by three to five people, one *Describer* and two or more *Seekers*. The Describer is shown an image, and assigns a descriptive caption which is broadcast to the Seekers. The Seekers search out the

[9] http://web.mit.edu/torralba/www/database.html

image from a database of images (e.g., the ESP dataset) and reveal it when they are confident. The correct Seeker then becomes the Describer. Phetch also extends the ESP model by introducing penalties for poor descriptions (from the Describer) and selecting the wrong image (from the Seeker). It also uses automated bots simulating players replaying decision to other players to improve the quality assurance on captions and images.

Magic Bullet - Yan and Yu (2009)
Magic bullet is an ESP inspired game consisting of two competing teams of two players each. As in the ESP game, team players are chosen and assigned to teams at random. Players are to agree on the textual meaning of a segmented CAPTCHA image. The first team to agree gets the score. Magic Bullet yielded as high as 98% labelling accuracy in one of their test studies.

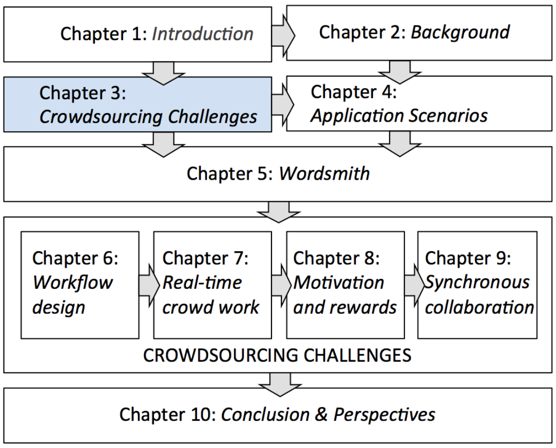
TagCaptcha - Morrison et al. (2009)
TagCaptcha is another CAPTCHA based, image annotation tool. However, unlike Magic Bullet, the images are not textual images but object images. TagCaptcha works in single actor mode wherein, a player is shown an image which is meant to have a specific one word label. The evaluation reported accuracy scores of 70% with players giving high level conceptual labels to images (e.g., ‘animal’ instead of ‘bear’ or ‘dear’).

SeaFish - Thaler et al. (2011)
SeaFish is a single player game in which a player is to select images that are semantically related to a concept (represented as an image from DBpedia). For example, distinguishing between an image of a *blackberry phone* and the *blackberry fruit*. Unlike the ESP type games, players do not type in actual labels, however, semantic data is generated from game play.

MOLT - Mavandadi et al. (2012)
Several studies have been carried out in applying crowdsourcing to the field of medical imaging. Mavandadi et al. (2012) presented *MOLT*, an interactive image labelling game where players are to identify malaria infected red blood cells. Their results reveal diagnostic accuracy scores within 1.25% of those by an expert medical professional.

MalariaSpot - Luengo-Oroz et al. (2012)
MalariaSpot is a game where players counted the number of parasites in blood smears to identify cases of malaria. The results yielded over 99% accuracy over 12,000 game plays in comparison with expert microscopists, which on the average spend 20 minutes identifying a single case.

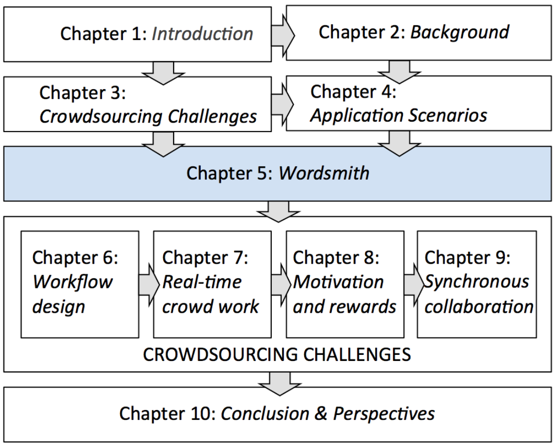
4.5 Summary


Chapter 1: *Introduction*
Chapter 2: *Background*
Chapter 3: *Crowdsourcing Challenges*
Chapter 4: *Application Scenarios*
Chapter 5: *Wordsmith*
Chapter 6: *Workflow design*
Chapter 7: *Real-time crowd work*
Chapter 8: *Motivation and rewards*
Chapter 9: *Synchronous collaboration*
CROWDSOURCING CHALLENGES
Chapter 10: *Conclusion & Perspectives*

In this chapter, we explored two broad application areas which form the basis of all our experiments in future chapters: text annotation and image labelling. We also discussed scenarios within each application area, before drawing a parallel with our four crowdsourcing challenges. Subsequently, we presented a literature review highlighting several bespoke crowdsourcing platforms which find relevance within each of the two broad application areas. This serves as a precursor to the next chapter on Wordsmith, our own custom-built crowdsourcing system.

Chapter 5

Wordsmith


Chapter 1: Introduction
Chapter 2: Background
Chapter 3: Crowdsourcing Challenges
Chapter 4: Application Scenarios
Chapter 5: Wordsmith
Chapter 6: Workflow design
Chapter 7: Real-time crowd work
Chapter 8: Motivation and rewards
Chapter 9: Synchronous collaboration
CROWDSOURCING CHALLENGES
Chapter 10: Conclusion & Perspectives

In this chapter, we introduce Wordsmith – our gamified platform for carrying out paid microtask crowdsourcing. We describe its interface design vis-à-vis each of the four challenges which we set out to address and the two application scenarios; next we describe our crowdsourcing process and how Wordsmith is used to carry out tasks from project definition to execution. Finally, we highlight how we improve submission quality by tackling malicious workers and evaluating worker submissions.

5.1 Introduction

Crowdsourcing provides a framework for leveraging on the scale and wisdom of the crowd to carry out tasks quickly and cost effectively. Paid microtask crowdsourcing operates via marketplaces where task requesters post tasks and potential workers search for and solve available tasks. This is different from the GWAP or Citizen Science models, which involve no financial remunerations. Several challenges arise in deploying paid micro tasks: from the technical to the ethical. This thesis is an attempt to study some of these challenges in-depth and to understand specific challenges associated with paid microtask crowdsourcing. To test our theories, we needed a flexible platform that was adaptable to our varying needs. We created *Wordsmith* not as 'yet another annotation tool', but as a system for carrying out experiments in paid microtask crowdsourcing where we had full control not only of the interface, but of all the incentives that sit on top of the base financial payments. Wordsmith overcomes limitations of conventional crowdsourcing marketplaces (such as CrowdFlower and Mechanical Turk) in the power it affords to design complex gamification interfaces and workflows. It also leverages on the marketplaces as a large source of available workers, and a medium for worker compensation.

We implemented Wordsmith to test out theories on addressing four of the twelve challenges posited by Kittur et al. (2013) in the context of gamified paid microtask crowdsourcing. We selected two of the most popular types of crowd task types as reported by Difallah et al. (2015): text annotation (or specifically Twitter annotation) and image labelling. These two task types were discussed earlier as crowdsourcing application scenarios in Chapter 4. In this chapter we present Wordsmith in detail, elaborating design choices and implementation strategies in line with the four challenges and two application areas.

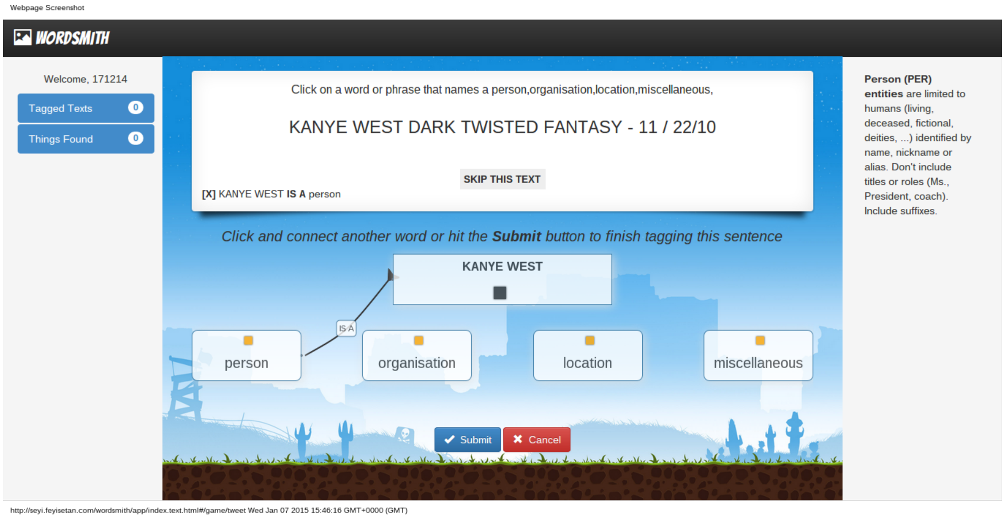
We also look at the annotation process and how Wordsmith supports each stage of the crowdsourcing lifecycle from project definition to data evaluation and aggregation. Since the specific interface design and the underlying design principle varies depending on the task mode, we leave the in-depth discussion for the related chapters. Figure 5.3 below shows the Wordsmith interface for labelling images.

5.2 Interface Design

Wordsmith is a gamified platform for carrying out paid microtasks. Unlike most of the other games surveyed in Chapter 4, Wordsmith sources its players (or workers) from crowdsourcing marketplaces such as CrowdFlower and Amazon Mechanical Turk. Wordsmith is used to carry out named entity recognition (NER) tasks on microposts. In NER tasks, workers are required to identify instances of person (PER), organisation (ORG), location (LOC) and miscellaneous (MISC) entities in tweets. Wordsmith is also used to carry out image labelling tasks - albeit at a much higher level (i.e., we do not carry out positional analysis or image segmentation). Wordsmith's interface features a three pane layout: the first pane to the left holds the annotator information and game state; the second (middle) pane holds the task details and user controls; while the last pane holds additional information, current leaderboard and global information. This is the general layout, however, the specific design varies based on the application scenario and the specific challenge being addressed. In the following sections, we briefly describe the modifications we made to address the four challenges. From Chapter 6, we describe the specific platform designs in more details.

5.2.1 Addressing: Workflow Design

In an earlier work, we described a way to quickly extract entities from large micropost datasets (Feyisetan et al., 2014). However, some of those tweets cannot be properly annotated automatically, while others need to be further disambiguated by a team of experts, hence we introduced a crowd component to carry out named entities on microposts. This version of Wordsmith (standard tweet annotation mode) sits as part of the design for a hybrid workflow for tweet annotating. As shown in Figure 5.1, workers are shown tweets with selectable words, which they can identify as instances of entity types. For example, in the Figure, a crowd worker could select the words 'Kanye' then 'West' and drag a descriptor to define 'Kanye West' as the instance of a person (PER). The task starts with a training phase after the workers have been sourced from CrowdFlower. In the training phase, the worker must annotate a tweet with known answers. Afterwards, the worker annotates a baseline set of tweets as requested by the task designer.


Webpage Screenshot

Welcome, 171214

Click on a word or phrase that names a person,organisation,location,miscellaneous,

KANYE WEST DARK TWISTED FANTASY - 11 / 22/10

[X] KANYE WEST **IS A** person

*Click and connect another word or hit the **Submit** button to finish tagging this sentence*

Person (PER) entities are limited to humans (living, deceased, fictional, deities, ...) identified by name, nickname or alias. Don't include titles or roles (Ms., President, coach). Include suffixes.

http://seyi.feyisetan.com/wordsmith/app/index.text.html#/game/tweet Wed Jan 07 2015 15:46:16 GMT+0000 (GMT)

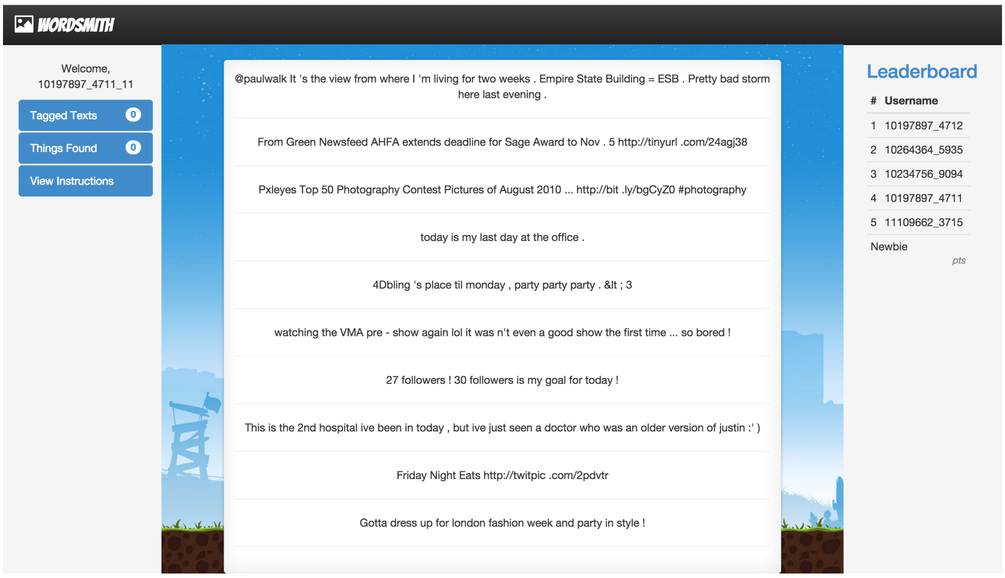
FIGURE 5.1: Wordsmith Interface Addressing: Workflow Design

The standard tweet annotator mode operates as a single actor platform with minimal gamification elements (no points, badges, levels or leaderboards). Workers have no perceivable knowledge of other users and are not given real-time scores based on their input. The system mirrors a traditional paid microtask platform, however, unlike the other systems, workers can skip tweets and select only the tasks they can confidently solve. Workers can also submit more tasks beyond the requested baseline. With these elements of choice, coupled with the annotation accuracy from the task output, we can get insight into what tweets the crowd is best at annotating. For example, if tweets mentioning organisations are continually and consistently skipped, while tweets referring to people are annotated accurately, then it implies workers are more suited to carrying out 'person' annotations. Piecing the information helps in designing effective workflows that combine automatic tools, inexperienced crowds and expert annotators.

The interface design and inner workings of the standard mode are detailed in Chapter 6.

5.2.2 Addressing: Real-time Crowd Work

To carry out real-time annotations, Wordsmith was modified to handle crowdsourcing contests. Wordsmith's contest mode for tweet labelling offers a more interactive and gamified outlook to named entity recognition when compared to the standard tweet annotation mode. It works as a multi-player competition for annotating named entities in tweets. The base interface and annotation layout is identical to the standard annotation mode. Workers can also skip tweets and annotate as many tweets as possible (within a fixed time constraint). However, unlike the standard mode, multiple workers connect to the platform concurrently to annotate a streaming set of tweets. Gamification elements are also present in the form of instant score computations based on the quality of individual worker annotations, and a leaderboard, which shows the relative rankings between connected workers. Figure 5.2 shows the contest layout to support real-time crowdwork. The middle pane holds the streaming tweets available for annotation for a set period of time (known as the *warping time* – more details in Chapter 7). When a worker selects any of the tweets, they are taken to an annotation interface identical to that displayed in Figure 5.1 where the tagging process follows the steps in the standard tweet annotation mode. The left pane displays the number of tweets annotated by the current worker; while the right pane holds the leaderboard and the number of currently connected competing workers.



@paulwalk It 's the view from where I 'm living for two weeks . Empire State Building = ESB . Pretty bad storm here last evening .

From Green Newsfeed AHFA extends deadline for Sage Award to Nov . 5 http://tinyurl .com/24agj38

Pxleyes Top 50 Photography Contest Pictures of August 2010 ... http://bit .ly/bgCyZ0 #photography

today is my last day at the office .

4Dbling 's place til monday , party party party . &lt ; 3

watching the VMA pre - show again lol it was n't even a good show the first time ... so bored !

27 followers ! 30 followers is my goal for today !

This is the 2nd hospital ive been in today , but ive just seen a doctor who was an older version of justin :')

Friday Night Eats http://twitpic .com/2pdvtr

Gotta dress up for london fashion week and party in style !

#	Username
1	10197897_4712
2	10264364_5935
3	10234756_9094
4	10197897_4711
5	11109662_3715
	Newbie
	pts

Figure 5.2: Wordsmith Interface Addressing: Real-time Crowd Work

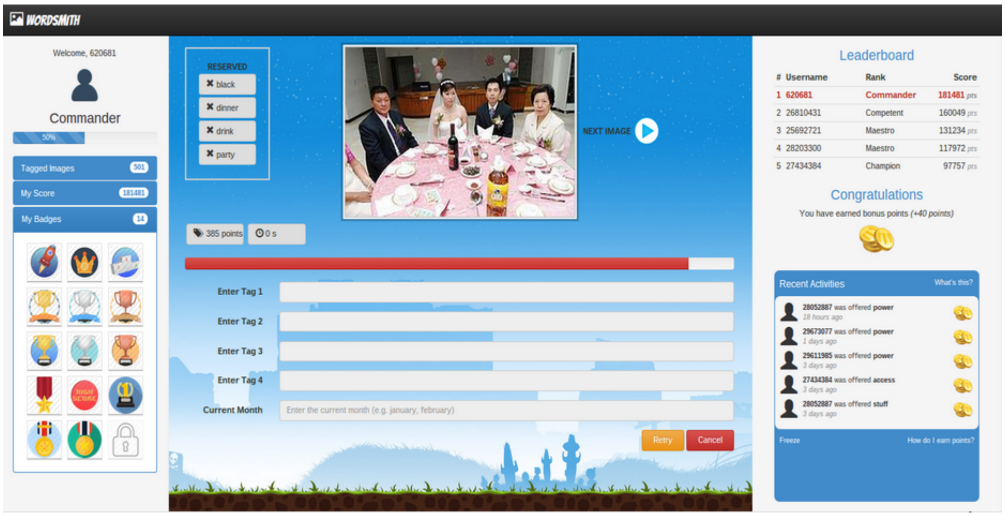
A fixed number of workers (default of 100) connect simultaneously to the platform. Workers can see the number of other contestants connected to the platform (updated constantly as some workers drop off the task). Workers select as many tweets as possible to annotate while each stream section is visible. The entire contest runs for about 6 minutes. Scores are awarded based on annotation agreements with an existing gold standard. The contest mode also contains a k-view leaderboard that is updated in real-time based on worker submissions. The leaderboard displays k workers ahead and behind the current worker's ranking. Scores are awarded for correct entity identification and typing so workers are incentivised to annotate correctly rather than try and make as many submissions are possible. In the contest mode, only a certain proportion of annotators are remunerated. To achieve this, Wordsmith generates an *exit code*, which is only displayed to workers who are eligible for payment. Eligibility criteria vary depending on the experiment condition – however, its primarily a function of worker ranking and the reward spread (i.e., the number of workers to be paid). The contest mode proved to be an effective way to elicit annotation judgements in near real-time. The interface design and inner workings are detailed in Chapter 7.

5.2.3 Addressing: Motivation and Rewards

We use Wordsmith's image labelling capabilities to test out theories of task motivation and incentives in order to address the challenge of motivation and rewards in paid microtask crowdsourcing. Just as Wordsmith's tweet annotator worked in two different modes, with each mode designed around a specific challenge; Wordsmith's image labeller also operates in two modes: (a) Single player mode (detailed in Chapter 8) addressing the challenge of motivation and rewards; and (b) Two-player mode (detailed in Chapter 9) addressing the challenge of synchronous collaboration. Wordsmith's single player image annotator is a fully gamified tool for labelling images as shown in Figure 5.3. The left pane holds the workers badges, displays the current score and achieved game level. The middle pane displays the actual image to be annotated, a list of restricted words, and a free text field to enter descriptive labels. The right pane holds a hourly leaderboard and an activities widget displaying the achievements of all the participating players in real-time. A player connected to the platform is shown an image and is required to provide associated keywords describing the image. Players enter free text, which is quality checked for correctness (as a valid English keyword) and uniqueness (the keyword is not repeated on the current or recent image). At later stages of the game, a restricted list of keywords is presented, and this narrows the word-space of the player. Several gamification mechanics are employed to keep the players engaged - such as points, levels, badges and a global leaderboard. Scores are awarded for unique and valid keywords submitted, while bonus points are awarded for keywords that match submissions in the gold standard (we describe this as a quasi gold standard as the labels cannot be regarded as complete descriptions of the images). Players cannot interact with other players, however, the leaderboard and activities widget keeps players aware of the presence of other players – this also serves as a source of task motivation.

5.2.3.1 Game Design

The design was heavily borrowed from the ESP game, with variations described below. The basic elements consisted of an image frame and text fields for inputting keywords. We describe Wordsmith in terms of the four defining properties of games introduced by McGonigal (2011) in her book '*Reality is Broken*' as follows: its goal, rules, feedback mechanisms and participation.



Leaderboard

#	Username	Rank	Score
1	620681	Commander	181481 pts
2	26810431	Competent	160049 pts
3	25692721	Maestro	131234 pts
4	28203300	Maestro	117972 pts
5	27434384	Champion	97757 pts

Congratulations

You have earned bonus points *(+40 points)*

28052887 was offered **power**
18 hours ago

29673077 was offered **power**
1 days ago

29611985 was offered **power**
3 days ago

27434384 was offered **access**
3 days ago

28052887 was offered **stuff**
3 days ago

FIGURE 5.3: Wordsmith Interface Addressing: Motivation and Rewards

Game Goal

The goal of the game was to annotate as many images as possible (up to the maximum in the dataset) with descriptive keywords. In designing Wordsmith, we incorporated several elements to engage the player in achieving the goal. We added progress timers, progress bars and feed forward alerts.

1. Progress Timer - we added a colour coded slider which showed the amount of time left to tag each image. The slider went from blue to green, then orange and red as the player ran out of time.

2. Progress Bars - a progress bar was included below the player's level to indicate how close the player was to advance to the next level.

3. Feed-forward Alerts - when a player was close to attaining a new badge or level, asides the progress bars, a subtle alert appeared to keep the player engaged with the system.

Rules and Constraints

Due to the simplicity of Wordsmith as an image labelling game, the rules of Wordsmith merely consist of constraints designed to prevent cheating and input from spam-bots. The game elements adopted are summarised as follows:

1. English Checker - the game made a call to a web service to ensure that the keywords were genuine English words. This improved the quality of tags submitted.

2. Reserved Words - (also known as taboo words) as players advanced into new levels, a set of restricted words were assigned to each image. This limited the keywords the players could use to tag the image.

3. Duplicate Checker - given the potential of players inputting multiple English words like *cat, cat, cat*, we checked that the same word wasn't inputted more than once before the player submitted.

4. Spam Checker - we alternately asked players to type in the current day of the week and the current year of the month as a simple human check. This was required after tagging every 10 images. This was limited as such, so as not to discourage genuine players.

Feedback Mechanisms

Feedback consists of information provided to players on their progress and current standing in the game. Providing feedback has been shown to improve player retention and engagement by enhancing intrinsic feelings of accomplishment as players advance. Some of the interface elements of Wordsmith are shown in Figure 5.4 below.

Voluntary Participation

The final trait was to present the task within the game as what the player chose to do rather than what they were mandated to do. In this regard, Wordsmith supported player freedom in three ways;

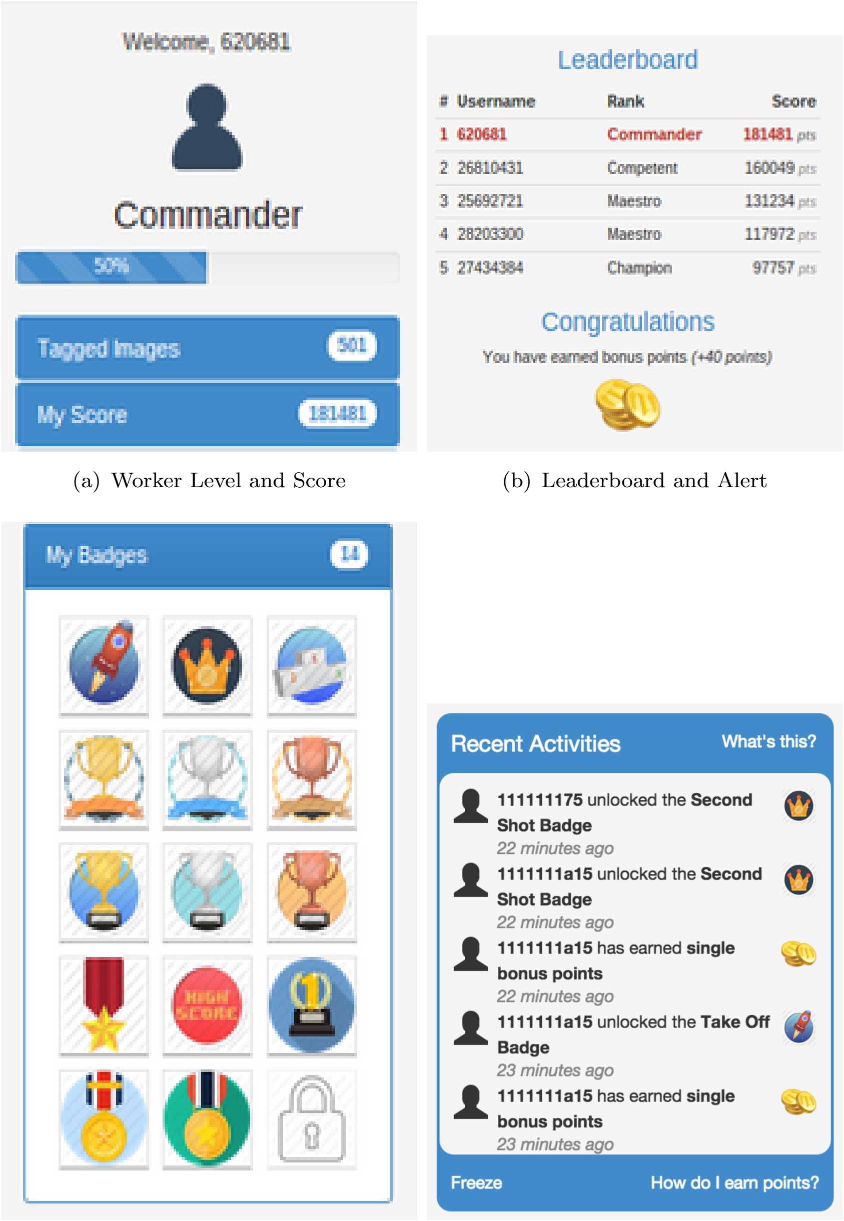
1. Optional Participation - players could simply join the game with a unique numeric ID. They did not need to fill out a registration form, provide any personal details or select a password.

2. Optional Exit - players could stop playing the game at anytime. In essence, players could actually tag less images than required or choose not to tag any images at all.

3. Optional Images - players could freely skip images they were not interested in and selectively tag images.

For all our image labelling experiments, we use the ESP dataset by von Ahn and Dabbish (2004) and source our players from CrowdFlower. The challenge being studied with the single player Wordsmith image labeller is the interplay between paid micro tasks and gamified platforms. We seek to understand worker motivations beyond financial payments and fun. We use the SAPS framework (Status, Access, Power and Stuff) to study the relationships between these high level sources of motivation. We also introduce

FIGURE 5.4: Wordsmith Interface Elements



(a) Worker Level and Score

#	Username	Rank	Score
1	620681	Commander	181481 pts
2	26810431	Competent	160049 pts
3	25692721	Maestro	131234 pts
4	28203300	Maestro	117972 pts
5	27434384	Champion	97757 pts

Congratulations

You have earned bonus points *(+40 points)*

(b) Leaderboard and Alert

(c) Worker Badges

111111175 unlocked the **Second Shot Badge**
22 minutes ago

1111111a15 unlocked the **Second Shot Badge**
22 minutes ago

1111111a15 has earned **single bonus points**
22 minutes ago

1111111a15 unlocked the **Take Off Badge**
23 minutes ago

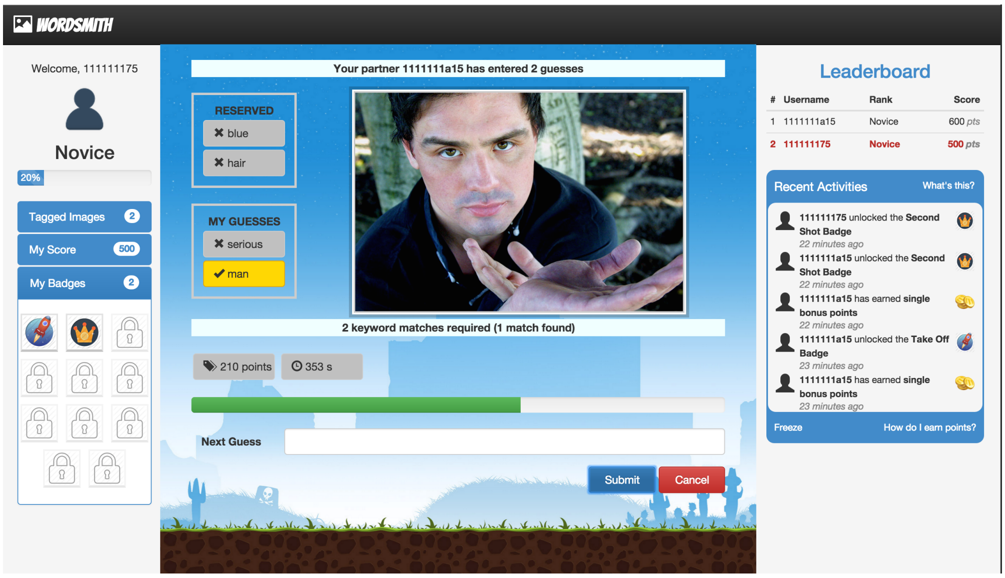
1111111a15 has earned **single bonus points**
23 minutes ago

(d) Activities Widget

and study the concept of *furtherance incentives* as a way to reduce the drop off rates from the game. Overall, the single player game mode represents an attempt to understand the challenge of keeping players engaged on paid microtask platforms. More details are presented in Chapter 8.

5.2.4 Addressing: Synchronous Collaboration

The Wordsmith two-player image annotator is more reminiscent of the original ESP game (von Ahn and Dabbish, 2004). A player connecting to the game is paired with another available player. Both players are shown the same image and are required to supply matching keywords to describe the image. The rules are quite similar to the single player mode, players must submit valid and unique keywords and might be restricted from using certain labels. Bonus points are also awarded for agreeing with labels in the gold standard dataset. The paired players cannot interact with each other and only receive feedback when they supply matching labels. Advancement in the task is collective (i.e., both players must match on a certain number of keywords to see the next image), however, scoring is individually based on the quality of submitted labels. The global leaderboard therefore reflected the individual scores of players (as the players could switch partners at any point in the game)



Your partner 1111111a15 has entered 2 guesses

2 keyword matches required (1 match found)

#	Username	Rank	Score
1	1111111a15	Novice	600 *pts*
2	111111175	Novice	500 *pts*

111111175 unlocked the **Second Shot Badge** *22 minutes ago*

1111111a15 unlocked the **Second Shot Badge** *22 minutes ago*

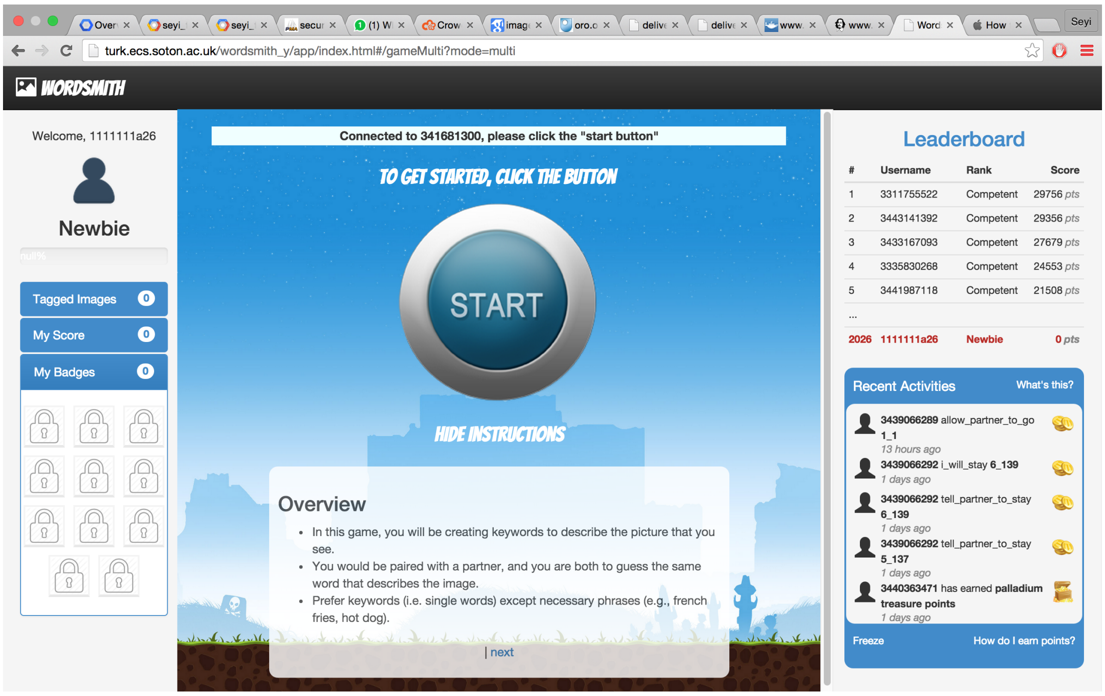
1111111a15 has earned **single bonus points** *22 minutes ago*

1111111a15 unlocked the **Take Off Badge** *23 minutes ago*

1111111a15 has earned **single bonus points** *23 minutes ago*

FIGURE 5.5: Wordsmith Interface Addressing: Synchronous Collaboration

The interface design is shown in Figure 5.5 and 5.6. It is almost identical to the single player annotator from Figure 5.3 sporting the same gamification elements (badges, levels, leaderboard and activities widget). In addition to these elements, there are additional notification areas in the middle pane to notify players of their connection to a partner player, as well as their partner's annotation activity. The two-player image annotator is not a new concept as it dates back to the original ESP game. The theory being understudied here is the impact of social pressure and social flow between the two collaborating players. As a paid gamified microtask platforms, players do not only partake in tasks for the fun of it (as with the ESP type games). However, players are required to complete a baseline set of micro tasks before they are eligible for payment. In the two player game however, since players can switch partners at any point in the game, they invariably become eligible for payments at different points in time. It is therefore inevitable for a player to leave another player who has not been paid in the middle of the game (however, Wordsmith would attempt to immediately reassign the unpaid player to a new partner). Our theory of social pressure is that the unpaid player can exert pressure to request the other player to remain in the game (and tag more



Connected to 341681300, please click the "start button"

Overview

- In this game, you will be creating keywords to describe the picture that you see.
- You would be paired with a partner, and you are both to guess the same word that describes the image.
- Prefer keywords (i.e. single words) except necessary phrases (e.g., french fries, hot dog).

Leaderboard

#	Username	Rank	Score
1	3311755522	Competent	29756 pts
2	3443141392	Competent	29356 pts
3	3433167093	Competent	27679 pts
4	3335830268	Competent	24553 pts
5	3441987118	Competent	21508 pts
...			
2026	1111111a26	Newbie	0 pts

Recent Activities

3439066289 allow_partner_to_go 1_1
13 hours ago

3439066292 i_will_stay 6_139
1 days ago

3439066292 tell_partner_to_stay 6_139
1 days ago

3439066292 tell_partner_to_stay 5_137
1 days ago

3440363471 has earned palladium treasure points
1 days ago

FIGURE 5.6: Wordsmith Collaboration Interface (newly paired players)

images than required) for the unpaid player to get paid. More details are presented in Chapter 9.

5.3 Crowdsourcing stages

Hovy and Lavid (2010) gave an overview of the annotation process. They identified seven questions that should be answered in the process of annotating corpora for NLP projects. This process is relevant for our tasks and we present and elaborate on them below.

- Selecting the corpus
- Instantiating the theory
- Selecting and training the annotators
- Specifying the annotation procedure
- Designing the annotation interface
- Choosing and applying the evaluation measures
- Delivering and maintaining the product

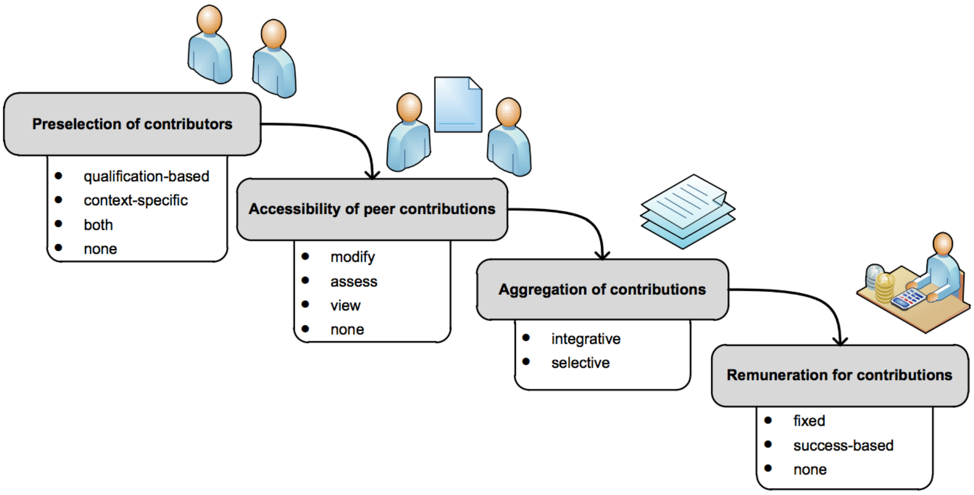
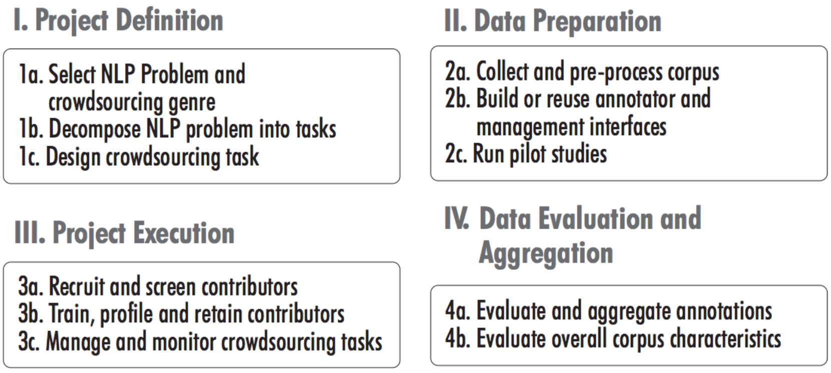

Preselection of contributors
qualification-based
context-specific
both
none
Accessibility of peer contributions
modify
assess
view
none
Aggregation of contributions
integrative
selective
Remuneration for contributions
fixed
success-based
none

FIGURE 5.7: Crowdsourcing process by Geiger et al. (2011)

Geiger et al. (2011) also presented a four step approach to crowdsourcing: preselection, accessibility, aggregation and remuneration. A guideline on corpus annotation through crowdsourcing was also presented by Sabou et al. (2014) wherein they highlight four stages from project definition through to data aggregation. These steps are shown in Figure 5.8. We now present how Wordsmith works through these four crowdsourcing stages and through the seven annotation process steps:



I. Project Definition

1a. Select NLP Problem and crowdsourcing genre
1b. Decompose NLP problem into tasks
1c. Design crowdsourcing task

II. Data Preparation

2a. Collect and pre-process corpus
2b. Build or reuse annotator and management interfaces
2c. Run pilot studies

III. Project Execution

3a. Recruit and screen contributors
3b. Train, profile and retain contributors
3c. Manage and monitor crowdsourcing tasks

IV. Data Evaluation and Aggregation

4a. Evaluate and aggregate annotations
4b. Evaluate overall corpus characteristics

FIGURE 5.8: Crowdsourcing stages (Sabou et al., 2014)

5.3.1 Project Definition

Instantiating the theory

Crowdsourcing in the research community begins with the instantiation of a theory. For example, in the later chapters of this work, we study challenges in incentives engineering, collaborative and real-time crowdsourcing. The associated theory determine the selected corpus (in the presence of several available corpora), task design, annotation categories, guidelines and procedures. Complex theories also lead to complex workflows resulting in iterative annotation instructions and pilot studies. Crowdsourcing task designs remain an art (as opposed to exact science), therefore, Wordsmith offers a flexible platform for technical designers to rapidly test and pivot on their initial theories.

Selecting the corpus

In instances where the theory can be studied on different datasets, selecting a suitable one becomes the next step. In some cases however, the corpus might be the subject of the theory (especially when it is the only one available). In cases of dataset multiplicity and availability, it is important to select one that is representative of the phenomenon to be studied in the theory. For example, in annotating language resources, the dataset should contain a natural distribution of words. The corpus creation date and sampling methods are also important as this can lead to variations in annotation results when the experiments are repeated at a later time (a related phenomenon of entity drift was reported by Fromreide et al. (2014)).

Specifying the annotation procedure

The annotation procedure is presented as instruction and guidelines for members of the crowd. In designing the procedure, the requester or team of experts go over the theory and task design, attempting to carry out the workflow on Wordsmith. This leads to instruction refinements such as: allowing for multiple option, increasing the context supporting the annotation decision or changing the sampling method or period of the dataset.

Once the guidelines are ready, Wordsmith provides interfaces for varying degrees of information policy adoption. A worker on engaging a task from a recruitment platform (e.g. CrowdFlower) sees a default set of instructions. Further in-line instructions can be given during task interaction - some of which are detailed in subsequent chapters. For example, in named entity recognition tasks, the right side bar is modified to hold additional information on the definition of entities and how to disambiguate them in one of our control experiment. In the image labelling tasks, additional guidelines come in the form of alerts that show how to attain points and why some labels are rejected.

Decomposing and designing the task

The requester determines: the number of workers; number of annotation categories and worker compensation. Wordsmith is able to programmatically limit the number of connections to one microtask, or set a minimum number of workers required for another. For named entity recognition tasks, the default is set to 3 workers. Wordsmith also supports utilising a variable number of workers, which sets it apart from traditional crowdsourcing platforms. This features prominently in tasks where Wordsmith is used as a GWAP platform given the differences in intrinsic worker motivation on these tasks.

This is also essential in choice based tasks where workers are encouraged to skip microtasks that they are not confident about. This can serve as a measure of task ambiguity, task difficulty or annotator confidence.

Wordsmith supports two broad classes of category numbering: (a) fixed - for example, in named entity recognition tasks where the entity types are pre-determined by the task requester; (b) variable - for example, in image labelling tasks, the range of allowed annotations is limited by the size of valid English words.

The task requester sets and manages the reward amount, however, we always advocate fair and ethical approaches to crowd work. Wordsmith does not handle reward payments as this is usually outsourced to the recruitment platform. Wordsmith however supports bonus strategies which link back to the worker source platform via an API call. Wordsmith provides options for different reward collection strategies. In the pure GWAP mode, workers can claim their payments at any point in time (even without completing the task). This is used for ascertaining fun as a primary motive for task engagement and not merely financial remuneration. In other modes, (for example, Wordsmith has support to be re-purposed as a contest platform), it might be essential to ensure task completion before payments are made. In these modes (e.g., the contest and collaborative modes), workers are issued an exit code on task completion. This is then entered into the recruitment platform which pays the worker.

5.3.2 Data Preparation

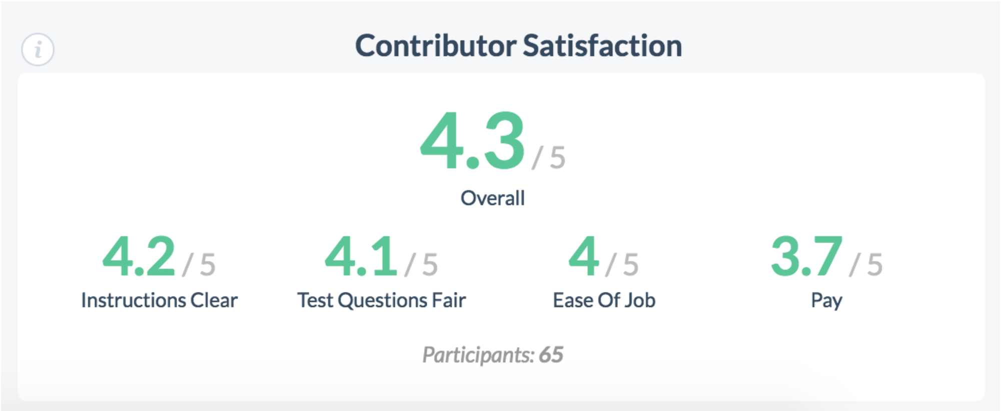
Corpus pre-processing
Wordsmith supports data integration in JSON and relational data formats. Wordsmith adopts a denormalised approach to data storage by ensuring the data is stored as closely to the required format with minimal number of joins. We run a set of pre-processing scripts before data is loaded into the datastore. This is usually task specific: for example, a task requester might need to process only English tweets, or strip out urls, usernames and #hashtags from microposts. Others might involve excluding certain images from an image labelling task based on a set of keywords that might render the image unsuitable for workplace annotation. Other pre-processing steps involve removing duplicates, resolving character-encoding issues, cleaning out blanks and setting a sampling mechanism for tasks which require a subset of the available dataset.

Gold standard creation
Wordsmith supports gold standard creation by simulating the base task at hand to be performed by a set of experts. To create the gold standard for entity types on tweets, a group of experts is given the task of annotating the tweets on Wordsmith. In performing this base task, Wordsmith is stripped of extraneous interface elements such as gamification mechanisms and time constraints. The annotated result is then

reformatted and copied over into the gold standard table while the original result set is emptied. It is also possible to import a preset gold standard, or a quasi gold standard (as in the case of image labels) to Wordsmith.

Pilot studies

Pilot studies are run on Wordsmith by experts, task designers and initial crowd workers to test the platform performance and tweak undecided variables. Much understanding of crowdsourcing has evolved as an art, and sometimes, it is necessary by trial and error to determine what works and what doesn't. Pilot studies utilise the full task design (as opposed to the stripped down base task used for gold standard creation) to understand how a larger number of crowd workers might interact with the full task. The task designer can set parameters to exclude and prevent pilot workers from the final study. Figure 5.9 shows worker satisfaction ratings in a pilot study.



Contributor Satisfaction

4.3 / 5
Overall

4.2 / 5
Instructions Clear

4.1 / 5
Test Questions Fair

4 / 5
Ease Of Job

3.7 / 5
Pay

Participants: 65

FIGURE 5.9: Worker satisfaction in pilot study

5.3.3 Project Execution

Worker recruitment

Wordsmith does not source for nor manage the payment of crowd workers, but only serves as a platform for advanced low level task performance. Crowd workers are recruited from external micro task marketplaces such as CrowdFlower, and then redirected to Wordsmith where they carry out the task, returning to the originating platform to receive their compensation. Most crowdsourcing marketplaces afford for high-level worker filters, for example, people who speak a certain language or from a certain country. Wordsmith does not store personal identifying information and as a result, is not able to carry out these filters.

Worker profiling

Wordsmith allows for worker profiling for various requester reasons. For example, a worker might be carrying out a within-subjects or between-subjects experiment which requires a set of workers be prevented from undertaking certain experiment conditions.

Other workers could be excluded because they were part of an initial pilot study, or the requester is rerunning an experiment due to either a corrupted set of results or for validation purposes. Malicious workers could also be targeted and blocked, and in some cases, it might be necessary to exclude power users from creating skewed results - although this is achieved by various other strategies such as using hourly leaderboards or post-processing to present a more balanced view. Profiling is done by using the worker IP address and the recruitment platform id.

Worker training

In crowdsourcing microtasks, workers generally have a small worldview of the overall aim of the task requester. As a result, worker training is essential to improve result quality and prevent unintentional errors. Worker training is also important to dispel underlying biases as to what the task requires - for example in pilot named entity annotation tasks without training, workers always tended to label *house* and *room* as named locations. In addition to the initial instructions and guidelines from the task recruitment platform, Wordsmith also displays additional instructions before the task and inline at a side bar during the task (this is configurable and could be removed to test for experiment conditions). By default in Wordsmith, the first worker task is always a training task with a known ground truth response. The task requester can also have know questions repeated at further instances of the task (the second default is halfway through the task). Wordsmith also features multiple non obtrusive alerts which do not prevent task continuance: (a) feedback alerts which inform workers how they achieved a high score or bonus point; (b) feedforward alerts which inform workers when they are close to a new level, or badge; (c) general alerts which notify on wrong spellings, duplicate entries or restricted entries.

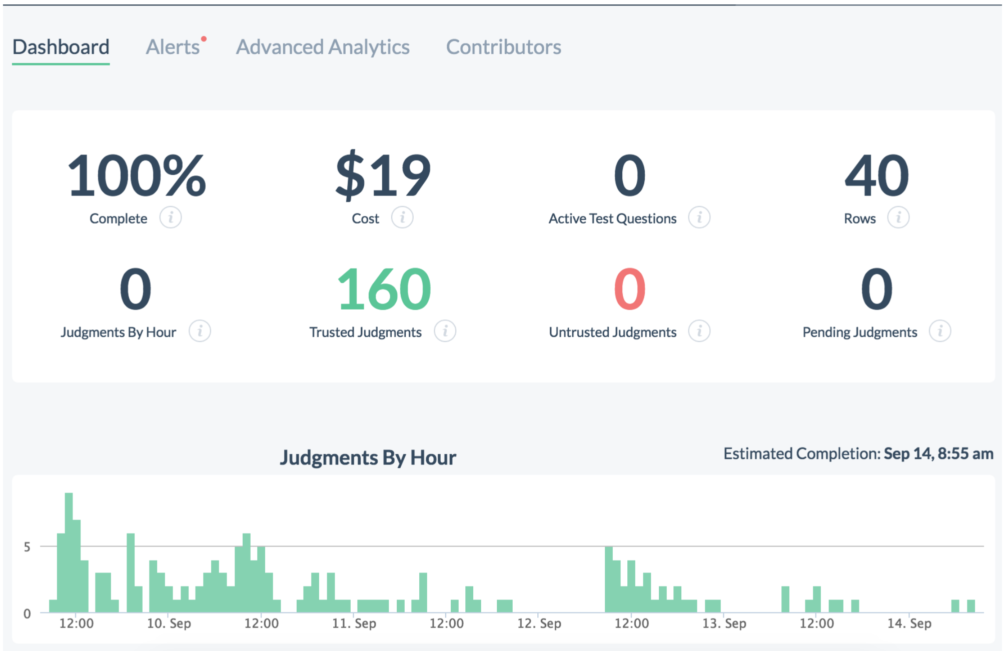
Worker scoring

Wordsmith's scoring philosophy is related to that of Chamberlain et al. (2008). As with other gamified systems as described by Zichermann and Cunningham (2011), points serve as a feedback mechanism to keep game players (or workers) aware of their task progress and relative positioning. The worker's first correct task is met with a reward (such as a badge) and a feedback alert on bonus points and valid answers. This serves as an on boarding mechanism while the worker gets conversant with the task at hand. After the initial task levels (e.g., newbie and novice in the image-labelling task), the point system is geared to motivate workers to produce better output by using a more intricate system of points based on the quality of their results with respect to an existing gold standard. Wordsmith affords for multiple scores in the form of bonus points and treasure points which can be configured by the task requester dependent on either hard or approximate matches with the gold standard.

The workers score represent not only their individual progress, but serves as a measure of how they stack up against others. Leaderboards are a mechanism to show such relative positioning. However, leaderboards can also demotivate new workers from performing beyond their required output if the scores of others appear unattainable. As a result, Wordsmith incorporates 3 leaderboard strategies: (a) a global leaderboard which shows the high scores of all times; (b) an hourly leaderboard which shows the top scores for the hour; and (c) a k-view leaderboard which shows (as the configurable default), 3 workers ahead of and 3 workers behind the current worker.

Wordsmith does not include negative scoring nor a zero point system. We sought other approaches to ensure quality control (detailed in section 5.4) as multiple pilot studies revealed a significant drop out of workers and multiple negative reviews in discussion forums.

Task management
Wordsmith includes a command line interface for task management and monitoring. This gives real time access to worker connections and task submissions. Figure 5.10 shows CrowdFlower's monitoring dashboard for an ongoing crowdsourcing task.



100% Complete

$19 Cost

0 Active Test Questions

40 Rows

0 Judgments By Hour

160 Trusted Judgments

0 Untrusted Judgments

0 Pending Judgments

Judgments By Hour

Estimated Completion: **Sep 14, 8:55 am**

Figure 5.10: CrowdFlower Task Monitoring Dashboard

5.4 Wordsmith Quality Control

5.4.1 Training and Evaluating Workers

The first line of defence in ensuring quality task submissions is by ensuring some form of training before task submission starts. As detailed in subsection 5.3.3, Wordsmith's worker training comes in form of instructions, guidelines and gold standard questions.

5.4.2 Worker Mistakes

Not every worker error is a malicious intent to subvert the task system. As a result of this understanding, Wordsmith incorporates several mechanisms to detect unintentional errors such as: (a) misspellings - in labelling tasks, Wordsmith connects to a dictionary web service to validate text input and notify workers of wrong spellings; (b) duplicates - workers are prevented from applying the same label more than once to a task and a further number of tasks as defined by the requester; and (c) restrictions -

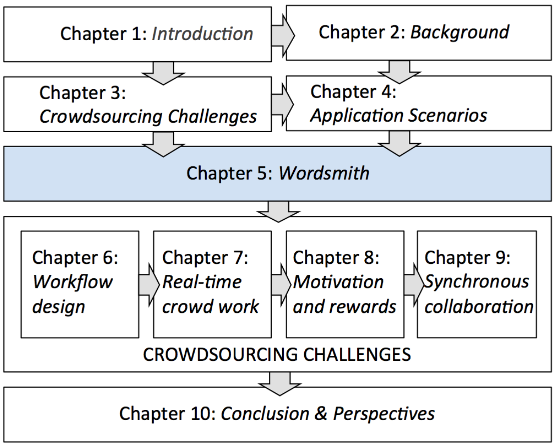
5.4.3 Malicious Workers

An essential part of attaining and maintaining high quality task output is by preventing malicious workers from participating. One of the core motivations of workers on paid microtask platforms is to receive financial compensation for their time. The more tasks they can complete, the more the potential financial reward. This can lead to mechanised submissions for tasks with simple workflows (e.g. always selecting the first option or typing in the same label). Wordsmiths worker profiling system (presented in section 5.3.3 above) helps to filter out workers based on their submission patterns. The profiling system does not automatically ban suspected input, rather, the alert system notifies workers about possible mistakes they could have made such as misspellings, duplicate entries and repeated entries. However, a worker who is suspected of using a task bot, or script injection of subverting the task input system is marked and banned from further assignments.

5.4.4 Multiple Judgments

Wordsmith tasks allow for multiple judgement collections. Workers are also allowed to submit more tasks than the baseline requirement - leading to a potentially richer set of answers. Handling multiple submissions is important for majority voting and to observe patterns of player choice. This becomes a quality issue when multiple entries come from malicious workers in a bid to attain a high score.

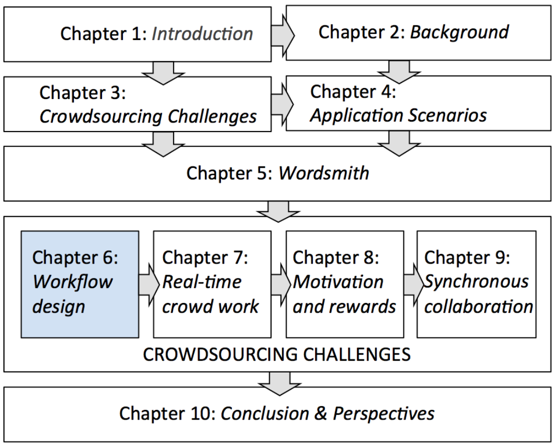
5.5 Summary


Chapter 1: *Introduction*
Chapter 2: *Background*
Chapter 3: *Crowdsourcing Challenges*
Chapter 4: *Application Scenarios*
Chapter 5: *Wordsmith*
Chapter 6: *Workflow design*
Chapter 7: *Real-time crowd work*
Chapter 8: *Motivation and rewards*
Chapter 9: *Synchronous collaboration*
CROWDSOURCING CHALLENGES
Chapter 10: *Conclusion & Perspectives*

In this chapter, we introduced Wordsmith – our gamified platform for carrying out paid microtask crowdsourcing. We described how its various gamified interface modes were used to address our four crowdsourcing challenges. Afterwards, we described our crowdsourcing process and how Wordsmith fits in from project definition to execution. Finally, we highlighted how we improve submission quality by tackling malicious workers and evaluating worker submissions.

Chapter 6

Workflow Design


Chapter 1: *Introduction*
Chapter 2: *Background*
Chapter 3: *Crowdsourcing Challenges*
Chapter 4: *Application Scenarios*
Chapter 5: *Wordsmith*
Chapter 6: *Workflow design*
Chapter 7: *Real-time crowd work*
Chapter 8: *Motivation and rewards*
Chapter 9: *Synchronous collaboration*
CROWDSOURCING CHALLENGES
Chapter 10: *Conclusion & Perspectives*

In this chapter we describe the methods, experimental set-up, and data used to address the challenge of designing a useful workflow for crowdsourcing named entities. We discuss the potential of building better workflows for paid microtasks by leveraging on insights into task features and worker preferences. We then present our results based on the experiments conducted, and summarise our core findings. This chapter also introduces the concept of furtherance incentives which is expanded in later chapters. We conclude with an overview of our contributions and an outline for future work.

This chapter is adapted from earlier published work [1] titled 'Towards Hybrid NER: A Study of Content and Crowdsourcing-Related Performance Factors'.

6.1 Overview

In our work reported in this chapter, we posit that just as certain textual features (such as proper syntax and sufficient context) make tweets amenable to automatic NER, certain features also lead to higher quality named entity annotation by crowd workers. This leads to the design of more advanced workflows as illustrated in Figure 6.1 (as opposed to the simplistic workflow earlier presented in Figure 3.2) where the initial processing

[1] This chapter is adapted from work that appeared at ESWC 2015 Feyisetan et al. (2015a)

divides tweets between automatic tools and the crowd, and subsequently between the crowd and experts.

6.2 Model

This chapter offers an in-depth study of the factors which influence the performance of the crowd in hybrid NER approaches for microposts. We categorise these feature factors in 2 broad classes:

1. **Content features** – inherent in the tweets such as number of entities, types of entities (such as persons, organisations, locations), character length of the tweet and the tweet sentiment; and
2. **Crowdsourcing features** – observed during annotation such as skipped true-positive posts, average time spent to complete the tasks, accuracy of the answers and the worker interaction with the user interface.

We analyse the impact of these features on the accuracy of the results, the timeliness of their delivery and their distribution in correct and incorrect annotations. In order to fully understand these factors, we also studied the importance of crowd annotation guidelines vis-à-vis the debate on the role of detailed guidelines as a means of improving human annotation (Aroyo and Welty, 2015).

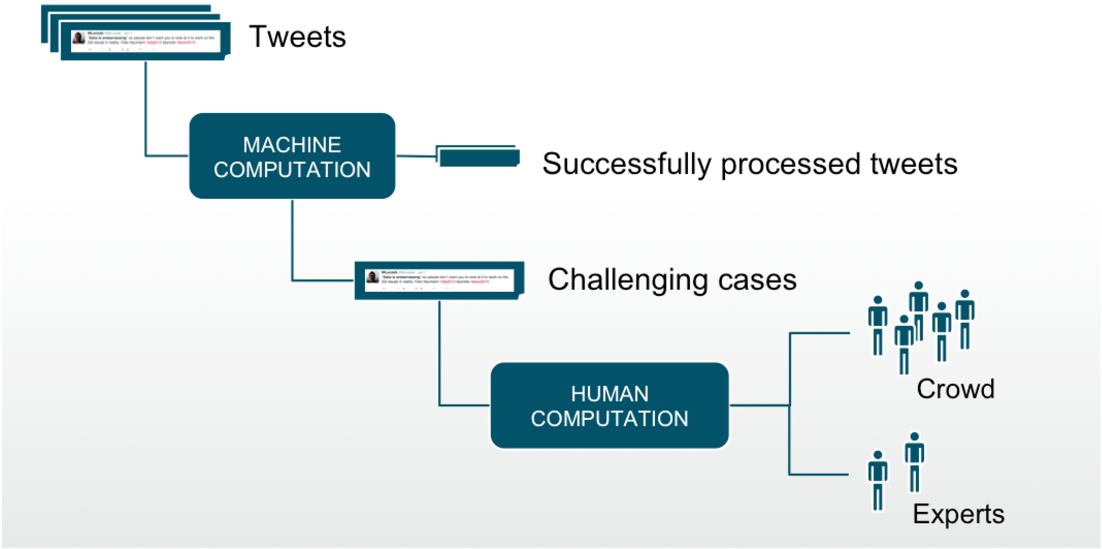


Figure 6.1: Proposed hybrid workflow

We run experiments on three datasets from related literature and a fourth newly annotated corpus using CrowdFlower and Wordsmith. An analysis of the overarching results reveal that detailed guidelines do not necessarily lead to higher quality annotations. The presence of additional disambiguating information however leads to specific annotation improvements such as annotating #hashtags and @mentions. Further analysis of the results illustrate that shorter tweets with fewer entities tend to be more amenable to microtask crowdsourcing. This applies in particular to those cases in which the text refers to single people or places, even more so when those entities have been subject to recent news or public debate on social media.

Though recommended by some crowdsourcing researchers and platforms, the use of the miscellaneous entity type as a NER category seems to confuse the contributors. However, it is well suited to identify a whole range of entities that were not explicitly targeted by the requester, from people who are less famous to partial, overlapping and what we call '*implicitly named entities*'.

6.3 Experiment Design

We used CrowdFlower to seek help from, select, and remunerate microtask workers; each CrowdFlower job included a link to our GWAP, which is where the NER tasks were carried out. Wordsmith was used to gather insight into the features that affect a worker's speed and accuracy in annotating microposts with named entities of four types: people, locations, organisations, and miscellaneous. The term 'GWAP' here is used lightly – as we did not design Wordsmith within the context of this study to include features which occur in traditional games (or gamified systems) such as badges, levels and activity widgets. Wordsmith however supports more bespoke functions which could not be easily achieved by using CrowdFlower.

6.3.1 Research Questions

Our basic assumption was that *particular types of microposts will be more amenable to crowdsourcing than others*; and that this insight can be used to design better crowdsourcing workflows to incorporate the crowd and experts. Based on this premise, we identified two related research hypotheses, for which we investigated two research questions:

[H1] Specific features of microposts affect the accuracy and speed of crowdsourced entity annotation.

RQ1.1. How do the following features impact the ability of non-expert crowd contributors to recognize entities in microposts:

- the number of entities in the micropost

- the type of entities in the microposts
- the length of micropost text
- the micropost sentiment

[H2.] We can evaluate crowd worker preferences for NER tasks.

RQ2.1. Can we evaluate crowd workers preferences for certain types of tasks by observing and measuring

- the number of skipped tweets (with entities that could have been annotated)
- the precision of answers
- the amount of time spent to complete the task
- the worker interface interaction (via a heatmap)

6.3.2 Research Data

We took three datasets from related literature, which were also reviewed by Derczynski et al. (2015). They evaluated automatic NER tools on these corpora, while we are evaluating crowd performance. The choice of datasets ensures that our findings apply to hybrid NER workflow, in which human and machine intelligence would be seamlessly integrated and only a subset of microposts would be subject to crowdsourcing. The key challenge in these scenarios is to optimize the overall performance by having an informed way to trade-off costs, delays in delivery, and non-deterministic (i.e., difficult to predict) human behaviour for an increase in accuracy. By using the same evaluation benchmarks we make sure we establish a baseline for comparison that allows us not only to learn more about the factors affecting crowd performance, but also about the best ways to combine human and machine capabilities. The three datasets are:

1. **The Ritter Corpus** by Ritter et al. (2011) which consists of 2, 400 tweets. The tweets were randomly sampled, however the sampling method and original dataset size are unknown. It is estimated that the tweets were harvested around September 2010 (given the publication date and information from Derczynski et al. (2015)). The dataset includes, but does not annotate Twitter *@usernames* which they argued were unambiguous and trivial to identify. The dataset consists of ten entity types.
2. **The Finin Corpus** by Finin et al. (2010) consists of 441 tweets which was the gold standard for a crowdsourcing annotation exercise. The dataset includes and annotates Twitter *@usernames*. The dataset annotates only 3 entity types: person, organisation and location. Miscellaneous entity types are not annotated. It is not stated how the corpus was created, however our investigation puts the corpus between August to September 2008.

3. **The MSM 2013 Corpus**, the Making Sense of Microposts 2013 Concept Extraction Challenge dataset by Basave et al. (2013), which includes training, test, and gold data; for our experiments we used the gold subset comprising 1450 tweets. The dataset does not include (and hence, does not annotate) Twitter *@usernames* and *#hashtags*. All four entity types (person, organisation, location and miscellaneous) are included in the dataset.

4. **The Wordsmith Corpus**, we also created and ran an experiment using our own dataset. In previous work of ours we reported on an approach for automatic extraction of named entities with Linked Data URIs on a set of 1.4 billion tweets (Feyisetan et al., 2014). From the entire corpus of six billion tweets, we sampled out $3,380$ English ones using *reservoir sampling*. This refers to a family of randomized algorithms for selecting samples of k items (e.g., 20 tweets per day) from a list S (or in our case, 169 days or 6 months from January 2014 to June 2014) of n items (for our dataset, over $30million$ tweets per day), where n is either a very large or an unknown number.

In creating the fourth gold standard corpus, we used the NERD ontology (Rizzo and Troncy, 2011) to create our annotations, e.g., a school and musical band are both sub-class-of **NERD:Organisation**, but a restaurant and museum, are sub-class-of **NERD:Location**.

The four datasets contain social media content from different time periods (2008, 2010, 2013, 2014) and have been created using varied selection and sampling methods, making the results highly susceptible to entity drift (Fromreide et al., 2014). Furthermore, all four used different entity classification schemes, which we normalized using the mappings from Derczynski et al. (2015). Table 6.1 characterizes the data sets along the features we hypothesize might influence crowdsourcing effectiveness.

Dataset overview				
Metric	Finin	Ritter	MSM2013	Wordsmith
Corpus size	441	2,400	1,450	3,380
Avg. Tweet length	98.84	102.05	88.82	97.56
Avg. @usernames	0.1746	0.5564	0.00	0.5467
Avg. #hashtags	0.0226	0.1942	0.00	0.2870
Avg. num of entities	1.54	1.62	1.47	1.72
No. PER entities	169	449	1,126	2,001
No. ORG entities	162	220	236	390
No. LOC entities	165	373	100	296
No. MISC entities	0	441	95	405
#hashtags annotated	NO	NO	NO	YES
@usernames annotated	YES	NO	NO	YES

Table 6.1: The four datasets used in our experiments

6.3.3 Experimental Conditions

We performed two experiments for each dataset; this means we evaluated 7,665 tweets.

Condition 1

For each tweet we asked the crowd to identify four types of entities (people, locations, organisations, and miscellaneous). We elicited answers from a total of 767 CrowdFlower workers, with three assignments to each task. Each CrowdFlower job referred the workers to a Wordsmith-based task consisting of multiple tweets to be annotated. Each job was awarded 0.05 USD to annotate at least 10 tweets with no bonus incentive. We will discuss these choices in Section 6.5. The workers were provided with annotation instructions detailing the various entity types and how to identify them. More details on the annotation guidelines are discussed in 6.5.2.

Condition 2

The second experiment condition built on the first with the same basic setup. For each tweet we asked the crowd to identify four types of entities (people, locations, organisations, and miscellaneous). Each CrowdFlower job referred the workers to a Wordsmith-based task consisting of multiple tweets to be annotated. Each job was awarded 0.05 USD to annotate at least 10 tweets with no bonus incentive. However, in the second condition, workers were presented with (i) more annotation instruction; (ii) entity type disambiguation instruction and (iii) an updated interface which presented the additional instructions before annotation and inline during annotation. Effectively, we sought to understand the impact more detailed instructions would have on worker accuracy (annotation speed, precision and recall).

We also carried out basic sentiment analysis on the tweet corpora, following in the steps of Saif et al. (2012) and Go et al. (2009). We hypothesized that particularly polarised tweets might have an effect on the entity annotation (Morris, 2011). For example, do workers annotate tweets with positive sentiments faster and more accurately compared to tweets about wars, outbreaks and tragedy. We used AlchemyAPI,[2] an external Web service providing natural language processing capabilities, in order to calculate the sentiment of each tweet to be annotated. AlchemyAPI was also used to carry out sentiment analysis on movie reviews from IMDb by Singh et al. (2013). Their results presented AlchemyAPI with an F1 score of 77.78% on a dataset of 1,000 reviews.

[2] AlchemyAPI – http:www.alchemyapi.com

6.3.4 Methods of Analyses

The outcome of the experiments was a set of tweets annotated with entities according to the four categories mentioned earlier. We measured the execution time and compared the accuracy of the crowd inputs against the four benchmarks. By using a number of descriptive statistics to analyse the accuracy of the users performing the task, we were able to compare the precision, recall and F1 scores for entities found within and between the four datasets. We also aggregated the performance of users in order to identify a number of distinguishing behavioural characteristics related to NER tasks. Our outcomes are discussed in light of existing studies in respects to the performance of the crowd and hybrid NER workflows. For each annotation, we measured data points based on mouse movements every 10 microseconds. Each point had an x and y coordinate value which was normalized based on the worker's screen resolution. These data points were used to generate the heatmaps for our user interface analysis. For each annotation, we also recorded the time between when the worker views the tweet to when the entity details are submitted.

6.4 Entity Types

We understood that the experiment settings would benefit from an harmonisation in the definitions of the entities. This is necessitated by the disparate nature of the entity type schemes used in the annotations of the different corpora. The entity type definitions from Finin et al. (2010) are as follows:

- *Person* (PER) - entities are limited to humans (living, deceased, fictional, deities, ...) identified by name, nickname or alias. This excludes titles or roles such as Mr., president or coach.
- *Organisation* (ORG) - entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure
- *Location* (LOC) - entities include names of politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments.
- *Miscellaneous* (MISC) - entities include other types of named entities, e.g., events (World War II), products (iPhone), animals (Cheetah), inanimate objects and monetary units (the Euro) etc.

6.4.1 Definitions and Mappings

We used the NERD ontology (Rizzo and Troncy, 2011) to normalise these definitions even though the results were slightly different from the entity mappings adopted by Derczynski et al. (2015). Our mappings assigned *musicartist* as person (PER), distinguishing it from *musicband* which we assigned as organisation (ORG). The gains in using the NERD ontology in spite of this slight mismatch meant we could have a reference baseline when dealing with more ambiguous cases e.g., organisation-location mismatches.

Entity Mappings				
Baseline	Finin	Ritter	MSM2013	Wordsmith
Person	person -	person musicartist	per -	person -
Organisation	org -	company sportsteam	org -	organisation musicalband
Location	place -	facility geo-loc	loc -	location -
Misc	-	movie product tvshow other	-	misc

Table 6.2: Entity mappings across the datasets

6.4.2 Difficult Cases

Organisation vs. location

In our preliminary experiments and gold standard creation, we noticed a number of cases that caused inter-annotator debate and disagreement. For example, given the tweets, *I am on my way to walmart* and *My local walmart made a lot of money last thanksgiving*, deciding the entity type of *Walmart* in context becomes difficult, even for expert annotators. This extends to other classes such as museums, restaurants, universities and shopping malls.

Organisation	Location
University	Museum
Education Institution	Restaurant
-	Shopping Mall
-	Hospital

Table 6.3: Adopted Organisation-Location Disambiguation

Software vs. organisation

We also noticed a number of tweets which mentioned software which were eponymous with their parent company. For example, '*Facebook bought the photo-sharing app, Instagram*' and '*I just posted a photo on facebook :)*'. The NERD ontology assigns pieces of software as a sub-class-of **NERD:Product** which maps to our miscellaneous (MISC) class. However, in cases such as these (*Facebook*, *Instagram*, *Google* and *Twitter*), we assign such entities as type organisation (ORG). For non-eponymous software or web applications e.g., *microsoft word, gmail*, these were mapped to the miscellaneous (MISC) class.

Typos, abbreviations and colloquialisms
Consider the tweet '*Road trip to see one of the JoBros' house w/ friends WHAT! WHAT!*'. The musical band Jonas Brothers has been replaced with a collapsed *urban* form. Other examples which underscore the difficulty of the task are tweets such as '*Marry jane is the baby tho*' where 'Mary' was misspelled as 'Marry' (which is another name for the psychoactive drug, marijuana). Similarly, '*Jack for Wednesday*', considering the capitalisation might refer to a footballer named Jack for the football club Sheffield Wednesday, or having Jack Daniel's whiskey for Wednesday night drinks.

Nested entities
Consists of entities which overlap and could potentially be annotated in multiple ways. For example, consider the following tweet from the Ritter corpus: *'Gotta dress up for london fashion week and party in style !'*. The correct entity in this case would be the event *london fashion week*, whereas, the workers might just annotate *London* as a location. This is also similar to identifying partial entity matches. For example, consider this tweet from the Wordsmith dataset *'Nice pass over New York City'*. The correct entity identifies New York City as opposed to a partial entity match targeting just New York.

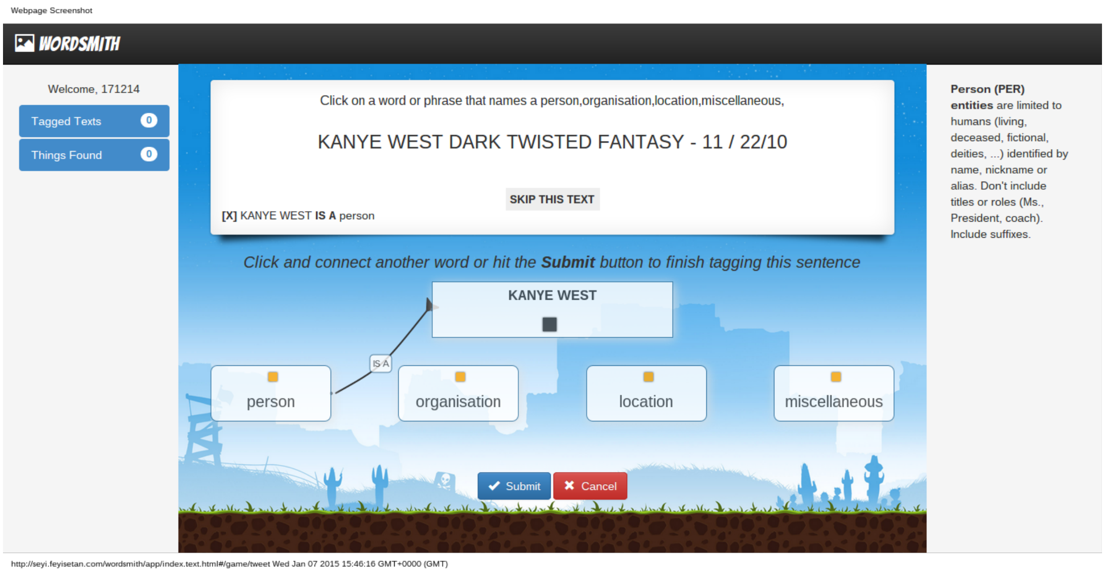
6.5 Crowdsourcing Approach

In this section, we would present an overview on our crowdsourcing approach. This includes details on our bespoke platform, our recruitment methodology using CrowdFlower, our reasons for not adopting a bonus system, our data and task model as well as our quality assurance strategy. We also elaborate on the annotation guidelines as it relates to the 2 experiment conditions, how we created our gold standard, and our approach to computing inter-annotator agreement scores.

6.5.1 Overview

Crowdsourcing platform: Wordsmith
As noted earlier, we developed a bespoke human computation platform called *Wordsmith* to crowdsource NER tasks. The platform is designed as a GWAP and sources

workers from CrowdFlower and has been discussed extensively in Chapter 5. A custom design approach was chosen in order to cater to an advanced entity recognition experience, which could not be obtained using CrowdFlower's default templates and markup language (CML). In addition, Wordsmith allowed us to set up and carry out the different experiments introduced in Section 7.5.


Webpage Screenshot

Welcome, 171214

Click on a word or phrase that names a person,organisation,location,miscellaneous,

KANYE WEST DARK TWISTED FANTASY - 11 / 22/10

[X] KANYE WEST **IS A** person

Click and connect another word or hit the ***Submit*** *button to finish tagging this sentence*

Person (PER) **entities** are limited to humans (living, deceased, fictional, deities, ...) identified by name, nickname or alias. Don't include titles or roles (Ms., President, coach). Include suffixes.

http://seyi.feyisetan.com/wordsmith/app/index.text.html#/game/tweet Wed Jan 07 2015 15:46:16 GMT+0000 (GMT)

FIGURE 6.2: Wordsmith interface

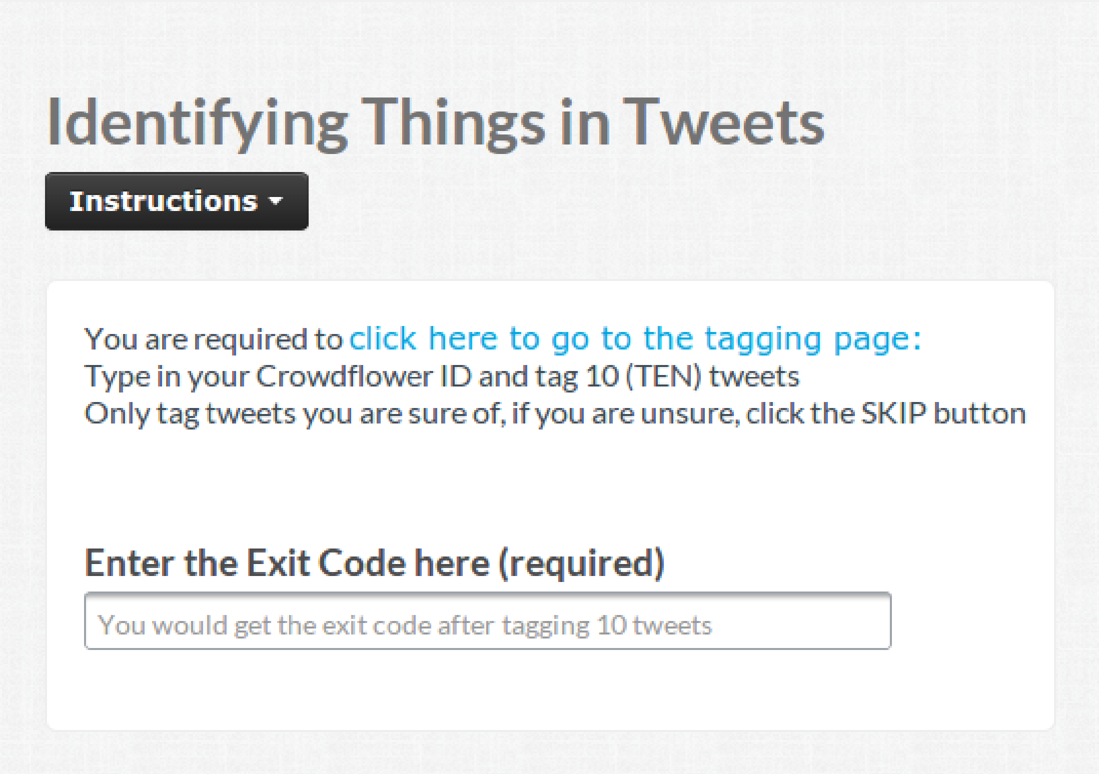
The tweet under consideration (as depicted in Figure 6.2) is presented at the top of the screen with each text token presented as a highlight-able span. The instruction to *'click on a word or phrase'* is positioned above the tweet, with the option to skip the current tweet below it. Custom interfaces in literature included radio buttons by Finin et al. (2010) and span selections by Braunschweig et al. (2013); Lawson et al. (2010); Voyer et al. (2010). We opted for a click-and-drag approach in order to fit all the annotation components on the screen, as opposed to Finin et al. (2010), and to cut down the extra type verification step by Braunschweig et al. (2013). By clicking on a tweet token(s) the user is presented with a list of connector elements representing the entity text and the entity types. Contextual information is provided in line to guide the user in making the connection to the appropriate entity type. When the type is selected, the type definition is displayed on the right hand side. The left sidebar gives an overview of the number of tweets the user has processed, and the total number of entities found. Once the worker has annotated 10 tweets, an *exit code* appears within the left side bar. This is a mechanism used to signal task completion in CrowdFlower, as we will explain in more detail later.

Recruitment

We sourced the workers for Wordsmith from CrowdFlower. Each worker was invited to engage with a task as seen in Figure 6.3, which redirected him/her to Wordsmith. After annotating 10 tweets via the game, the worker was presented with an exit code, which was used to complete the CrowdFlower job. We recruited *Level 2 contributors*, which are top contributors who account for 36% of all monthly judgements on the CrowdFlower platform (Feyisetan et al., 2015b). Since we were not using expert annotators, we set the judgement count at 3 answers per unit i.e., each tweet was annotated by three workers. Each worker could take on a single task unit; once starting annotating in Wordsmith, they were expected to look at 10 tweets to declare the task as completed. However, they were also allowed to skip tweets (i.e., leave them unannotated) or continue engaging with the game after they reached the minimum level of 10 tweets. Independently of the actual number of posts tagged with entities, once the worker had viewed 10 of them and received the exit code, he/she receives the reward of 0.05 $.

Bonus system

Unlike Lawson et al. (2010) or Yetisgen-Yildiz et al. (2010), we did not use any bonuses. The annotations carried out in Lawson et al. (2010) were on emails with an average length of 405.39 characters while the tweets across all our datasets had an average length of 98.24 characters. Workers in their case had the tendency to under-tag entities, a behaviour which necessitated the introduction of bonus compensations which were limited and based on a worker-agreed threshold. The tasks in Yetisgen-Yildiz et al. (2010) use biomedical text, which according to them, '[is] full of jargon, and finding the three entity types in such text can be difficult for non-expert annotators'. Thus, improving recall in these annotation tasks, as opposed to shortened and more familiar text, would warrant a bonus system.



Identifying Things in Tweets

You are required to click here to go to the tagging page:
Type in your Crowdflower ID and tag 10 (TEN) tweets
Only tag tweets you are sure of, if you are unsure, click the SKIP button

Enter the Exit Code here (required)

You would get the exit code after tagging 10 tweets

Figure 6.3: Crowdflower interface

Input data and task model

Each task unit refers to N tweets. Each tweet contains $x = \{0, ..., n\}$ entities. The worker's objective is to decide if the current tweet contains an entity and correctly annotate the tweet with their associated entity types. The entity types were person (PER), location (LOC), organisation (ORG), and miscellaneous (MISC). We chose our entity types based on the types mentioned in the literature of the associated datasets we used. Our task instructions encouraged workers to skip annotations they were not sure of. As we used Wordsmith as task interface, it was also possible for people to continue playing the game and contribute more, though this did not influence the payment. We report on models with adaptive rewards elsewhere (Feyisetan et al., 2015b); note that the focus here is not on incentives engineering, but on learning about content and crowd characteristics that impact performance. To assign the total set of $7,665$ tweets to tasks, we put them into random bins of 10 tweets, and each bin was completed by three workers.

Output data and quality assurance

Workers were allowed to skip tweets and each tweet was covered by one CrowdFlower job viewed by three workers. Hence, the resulting entity-annotated micropost corpus consisted of all $7,665$ tweets, each with at most three annotations referring to people, places, organisations, and miscellaneous. Each worker had two gold questions presented to them to assess their understanding of the task and their proficiency with the annotation interface. Each gold question tweet consisted of two of the entity types that were to be annotated. The first tweet was presented at the beginning, e.g., *'do you know that Barack Obama is the president of USA'* while the second tweet was presented after the worker had annotated five tweets, e.g., *'my iPhone was made by Apple'*. The workers are allowed to proceed only if they correctly annotate these two tweets. We display the second tweet at a fixed point in order to simplify our analysis and remove bias arising from workers viewing the tweet at random intervals.

6.5.2 Annotation Guidelines

In each task unit, workers were required to decide whether a tweet contained entities and annotate them accordingly. As a baseline for both experiment conditions, we adopted the annotation guidelines from Finin et al. (2010) for person (PER), organisation (ORG) and location (LOC) entity types. We also included a fourth miscellaneous (MISC) type, based on the guidelines from Ritter et al. (2011).

In computational linguistics, annotation guidelines present arbitrary and often debatable decisions (Plank et al., 2014) as seen from the varying choices in our experiment datasets. The decision to annotate (or not to) *#hashtags*, *@mentions* and MISC types represent the beginning of choices which extends to guidelines on specific entity types. Some authors have argued that more detailed guidelines do not improve annotation quality

Aroyo and Welty (2015); while some others skip the guidelines altogether when dealing with experts (Plank et al., 2014). The latter category relies on the experts to make adhoc consensual judgements amongst themselves to address hard cases.

In our study, we experimented with 2 guideline conditions to observe the results of varying the amount of annotation guidelines.

Experiment condition 1

Instructions were presented at the start of the CrowdFlower job via the Wordsmith interface and in-line during annotation. Whenever a worker is annotating a word (or phrase), the definition of the currently selected entity type is displayed in a side bar. These instructions included the following: the task title, stated as *Identifying Things in Tweets*; an overview on the definition of entities (with a few examples); a definition of the various entity types (PER, ORG, LOC, MISC), including examples of what constitutes and does not constitute inclusion into the type categories.

Experiment condition 2

In condition 2, we provided more instructions. This included the title, stated as *Identifying Named Things in Tweets* and details on ways to handle 7 special cases. The special cases were (i) disambiguating locations such as restaurants and museums; (ii) disambiguating organisations such as universities and sport teams; (iii) disambiguating musical bands; (iv) identifying eponymous software companies; (v) dealing with nested entities by identifying the longest entities; (vi) discarding implicit unnamed entities such as hair salon, the house, bus stop; (vii) identifying and annotating *#hashtags* and *@mentions*. These instructions were placed as in *Condition 1*, with the addition of an interface update, which allowed the workers to review the additional instructions during annotation.

6.5.3 Gold Standard Creation

The gold standard used for our Wordsmith dataset was curated by 3 expert annotators (PhD and Post Doctoral researchers within the field). We manually tagged the tweet entity types using the Wordsmith platform. The Wordsmith corpus consisted of $3,380$ tweets, sampled between January 2014 to June 2014. Each tweet was annotated with the 4 designated entity types (PER, ORG, LOC, MISC). Unlike the other 3 datasets, we chose to annotate *#hashtags*. This decision was partially motivated by the nature of the dataset which had a significant number of event based *#hashtags* corresponding to the FIFA World Cup. Similarly, unlike the Ritter and MSM2013 datasets, we also annotated the *@usernames*. Our annotation choices comprised of a separation of entity types such as musical artists and musical bands as person (PER) and organisations (ORG) respectively.

6.5.4 Inter-annotator Agreement

The inter-annotator agreement describes the degree of consensus and homogeneity in judgments among annotators (Nowak and Rüger, 2010) and is seen as a way to judge the reliability of annotated data (Ramanath et al., 2013). Setting an inter-annotator threshold can enhance the precision of results from the crowd. It can be further used to shed light on our research question about crowd worker preferences for NER tasks (H2 RQ 2.1). Various scores such as the Kappa introduced by Cohen in Cohen (1960) have been used to calculate inter-rater agreement.

The inter-annotator agreement (or degree of disagreement) can also serve as a measure of the difficulty of the task – and can draw light unto 'hard cases' which might require further attention (Plank et al., 2014) and (Aroyo and Welty, 2013). Annotator disagreement is not limited to crowd workers only but extends to experts also. The authors of Aroyo and Welty (2013) argue that inter-annotator disagreement is *not noise, but signal*; and, Plank et al. (2014) incorporates it in the loss function of a structured learned for parts of speech tagging and named entity recognition.

We use the approach by Bhowmick et al. (2008) to determine the pair-wise agreement on an annotated entity text and types. Given **I** as the number of tweets in a corpus, **K** is the total number of annotations for a tweet, **H** is the number of crowd workers that annotated the tweet and **S** is the set of all entity pairs with cardinality $|S| = \binom{K}{2}$, where $k_1 = k_2 \ \forall \ \{k_1, k_2\} \in S$.

Given a tweet $\boldsymbol{i}$ and an annotated entity $\boldsymbol{k}$ where $\{k, k\} \in S$, the average agreement, A_{ik}, on the keyword $\boldsymbol{k}$ for the tweet $\boldsymbol{i}$ is given by

$$A_{ik} = \frac{n_{ik}}{\binom{H}{2}} \tag{6.1}$$

where n_{ik} is the number of human agent pairs that agree that annotation $\boldsymbol{k}$ is in the tweet $\boldsymbol{i}$.

Therefore, for a given tweet $\boldsymbol{i}$ the average agreement over all assigned annotations is

$$A_i = \frac{1}{|S|\binom{H}{2}} \sum_{k \in S}^{S} n_{ik} \tag{6.2}$$

We presented the average inter-annotator agreement for each corpus in the experiment in Table 6.13. We also presented the change in precision and recall values based on the inter-annotator thresholds in Table 6.15.

	Condition 1: Worker annotations			Condition 2: Worker annotations		
Entity type	Precision	Recall	F1 score	Precision	Recall	F1 score
Finin dataset						
Person	**68.42**	58.96	**63.34**	43.65	**49.36**	46.33
Organisation	50.94	27.84	36.00	38.43	33.06	35.54
Location	66.14	**60.71**	63.31	**60.78**	47.67	**53.43**
Miscellaneous	-	-	-	-	-	-
Ritter dataset						
Person	42.93	**69.19**	52.98	32.68	**65.72**	43.65
Organisation	28.75	39.57	33.30	27.82	42.26	33.55
Location	**67.06**	50.07	**57.33**	**62.22**	51.42	**56.31**
Miscellaneous	20.04	20.23	20.13	16.06	22.98	18.91
MSM2013 dataset						
Person	**87.21**	**86.61**	**86.91**	**78.26**	**80.69**	**79.46**
Organisation	43.27	38.77	40.90	53.10	38.37	44.55
Location	60.57	67.29	63.75	49.35	59.47	53.94
Miscellaneous	10.44	29.11	15.37	5.98	30.11	9.98
Wordsmith dataset						
Person	**79.23**	71.41	**75.12**	**75.95**	57.90	**65.71**
Organisation	61.07	53.46	57.01	35.97	32.30	34.04
Location	72.01	**72.91**	71.26	63.34	**65.17**	64.24
Miscellaneous	27.07	47.43	34.47	8.03	19.37	11.35

Table 6.4: *Experiment results* - Precision and Recall on the four datasets.

6.6 Results

The following sections present an in-depth run through of the results from the two experiment conditions.

6.6.1 Overview

Overview of Annotations

Table 6.6 gives an overview into how workers performed at the tweet level across the various datasets. The results suggest consistently that workers correctly annotate tweets with fewer entities. This result was consistent across the four datasets. We did not see any strong connection between the length of the tweet and the likelihood of it being annotated correctly or incorrectly, as the differences were not significant. The length of the tweet however determines the whether the tweet would be selected for annotation or not - and we discuss this in detail in a later section.

Correct Annotations

The results of our experiment with condition 1 and 2 are summarised in Table 6.4. The

Dataset	Experiment Condition 1				Experiment Condition 2			
	Confusion matrix (vs gold)				Confusion matrix (vs gold)			
	PER	ORG	LOC	MISC	PER	ORG	LOC	MISC
Finin	78	1	7	-	498	25	67	-
	1	27	5	-	52	334	27	-
	1	4	84	-	2	56	431	-
	-	-	-	-	-	-	-	-
Ritter	765	7	26	20	2112	22	53	61
	10	140	62	88	51	503	120	204
	9	9	751	22	32	17	1265	30
	15	46	29	217	30	106	37	500
MSM2013	3,828	25	8	7	4259	78	4	10
	16	299	13	28	23	582	13	12
	13	21	321	5	9	23	267	8
	12	82	5	91	30	81	7	111
Wordsmith	5,230	34	29	32	1750	11	12	26
	93	811	30	46	50	200	21	36
	25	58	1,078	8	20	68	439	0
	50	113	12	718	218	48	13	102

Table 6.5: *Experiment results* - Confusion Matrix on the four datasets.

Correct and Incorrect Annotations				
Dataset	**Correct**		**Incorrect**	
	Num of Entities	Tweet length	Num of entities	Tweet length
Finin	1.17	91.63	1.48	92.53
Ritter	1.24	106.02	1.61	99.02
MSM	1.19	98.95	1.81	97.02
Wordsmith	1.38	97.88	1.70	96.10

Table 6.6: *Experiment results* - Correct and Incorrect Annotations

first set of results in Table 6.4 contains precision, recall and F1 values for the four entity types for all four datasets. The results in the 2 experiment conditions (C1 and C2) indicate the same result patterns with matching entity types yielding the top precision and recall values. The results also present an average decrease in precision, recall and F1 scores from C1 to C2. This is in spite of the additional annotation guidelines presented in C2. This result is in line with *Myth 3* presented by Aroyo and Welty (2015) which states that detailed guidelines do not always yield better annotation quality. The results reveal highest precision scores in identifying PER entities. The only exception to this was in the Ritter dataset where the highest precision scores were in identifying LOC entities. The highest recall scores were split in between PER entities in the Ritter and MSM datasets and LOC entities in the Finin and Wordsmith datasets. However, the margins were less than 2% with a higher score recorded for PER entities in the C2 for the Finin dataset.

Incorrect Annotations

Figure 6.4 illustrates the entity types which were wrongly annotated by workers. Across all the datasets, we observe that the ORG and MISC entity types were consistently wrongly annotated. This was the case across the four datasets. This suggests that workers had the greatest difficulties in either identifying these entity types, or were wrongly assigning them to other entity types. We therefore computed a confusion matrix to have a clearer insight into what entity types were wrongly annotated, and how they were wrongly annotated.

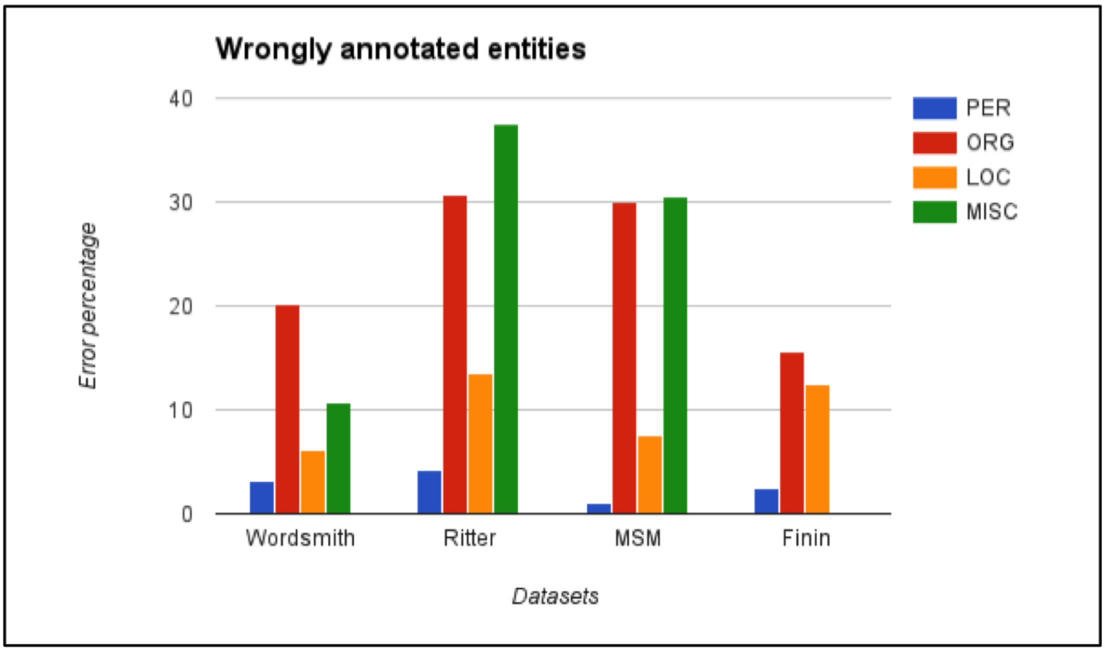


FIGURE 6.4: Incorrect annotations

Mismatched Annotations

We included a confusion matrix in Table 6.5 highlighting the entity mismatching types e.g., assigning *Cleveland* as location when it refers to the basketball team. The results suggest that the entity type ORG was mostly wrongly annotated as PER (in the Wordsmith dataset) and as MISC (in the Ritter dataset). The entity type LOC was most confused as the entity type ORG across all datasets (with the exception of the Ritter corpus). The typical confusion of the ORG and LOC types is a case of metonymy where these entities have to be especially handled in context (Maynard et al., 2003). This is seen where an organisation is associated with its location e.g., *Wall Street* and *Hollywood*. This phenomenon occurred in both experiment conditions even when more detailed instructions were given. In all dataset results, the MISC type was wrongly assigned the ORG entity type. The confusion matrix on the PER entity type was spread across all the other entity types. The Finin and Ritter showed the least confusion variance on the entity types across the two experiment conditions.

Skipped Tweets: Tweet Overview

Our guidelines encouraged workers to skip tweets for which they could not give confident annotations. Table 6.7 like Table 6.10 gives further insight into the dynamics of skipped tweets. The table presents, for C1 and C2, and across all datasets, the average number of entities present in a skipped tweet, as well as in an unskipped annotated tweet. The table also summarises, for both experiment conditions, and all datasets, the average number of characters in a skipped tweet and unskipped tweet. The tweets under consideration in the table are skipped true positive tweets i.e., tweets that were not annotated despite the presence of at least one entity.

The results highlight across all datasets, that workers skipped tweets that contained more entities than the ones they annotated on average. The results present evidence that workers on average skipped longer tweets. The results were consistent across the four datasets and between the two experiment conditions. The tweet length was least significant in the MSM2013 experiment (with the number of characters between the skipped and unskipped tweet differing by less than 1 character), once again due to the comparatively well-formed nature of the dataset and the least standard deviation in the tweet lengths. The tweet length feature was most significant in the Ritter dataset, with workers systematically skipping tweets that were significantly longer than the average tweet length; it is worth mentioning that this corpus comprised the highest average number of characters per micropost.

We do not report a high level metric on the number of tweets skipped, as this might have been misleading. For example, given 10 tweets annotated by 3 workers, the tweets skipped by each worker might have been annotated by another. We therefore present fine-grained results on the distribution of entity types present in tweets skipped by individual workers and the tweet sentiment. We also report aggregate findings on the average number of entities present in, and the average length of skipped tweets

Skipped Tweets: Entity Types

More results on the skipped true-positive tweets are presented in Table 6.8 and Figure 6.5. It contains the distribution of the entities present in the posts that were left unannotated in each dataset according to the gold standard. On average across all four datasets, people tend to avoid recognizing organisations, but were more keen in identifying locations. In the MSM2013 dataset, person entities were least skipped due to the features of the dataset discussed earlier (e.g., clear text definition, consistent capitalisation etc.). The entity types in the Wordsmith dataset (apart from the LOC type) were all skipped with equal likelihoods.

Condition 1: Skipped tweets				
Dataset	**Skipped**		**Annotated**	
	Num of Entities	Tweet length	Num of entities	Tweet length
Finin	1.56	101.39	1.33	94.82
Ritter	1.42	113.05	1.35	104.22
MSM	1.49	98.74	1.30	97.11
Wordsmith	1.62	102.22	1.39	97.84
Condition 2: Skipped tweets				
Dataset	**Skipped**		**Annotated**	
	Num of Entities	Tweet length	Num of entities	Tweet length
Finin	1.51	102.44	1.20	98.99
Ritter	1.52	112.08	1.00	104.68
MSM	1.50	100.4	1.23	99.51
Wordsmith	1.61	102.70	1.39	98.14

Table 6.7: *Experiment results* - Skipped true-positive tweets

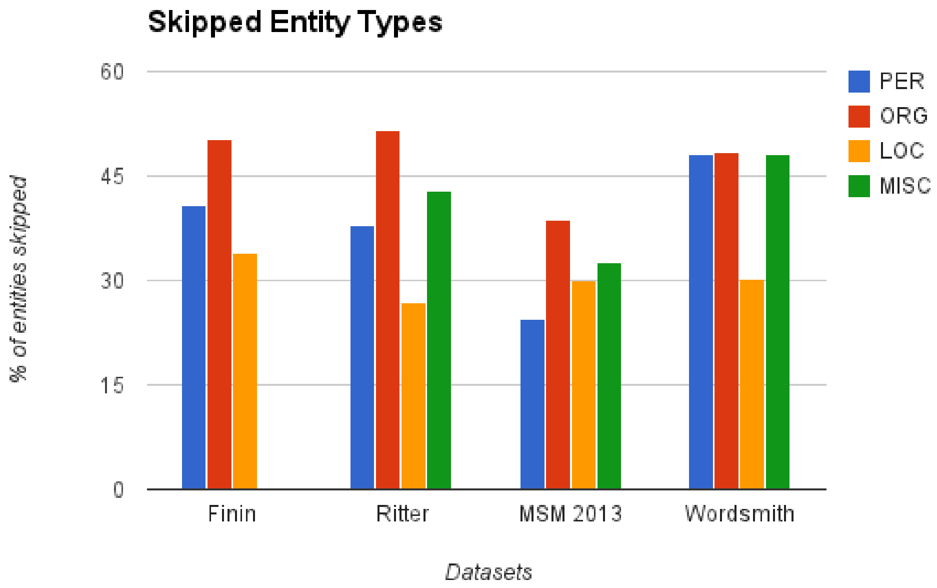


Figure 6.5: *Skipped Tweets*: Entity Types in Skipped Tweets

We posit this to be as a result of two factors: our uniform sampling method which did not bias the presence of a single entity type (e.g., as in the MSM2013 dataset) and increased use of *@mentions* and *#hashtags* in the dataset. This result is also in line with those presented in Table 6.5 that ORG was the most misidentified entity type. This result was consistent across both experiment conditions with crowd workers still skipping tweets with organisation entities when more instructions were given on how to disambiguate them.

Condition 1: Skipped true-positive tweets				
Dataset	PER	ORG	LOC	MISC
Finin	40.91% (90/220)	50.27% (93/185)	33.83% (68/201)	-
Ritter	38.01% (631/1660)	51.57% (361/700)	26.83% (501/1867)	42.95% (847/1972)
MSM 2013	24.35% (1200/4928)	38.81% (437/1126)	30.13% (185/614)	32.58% (129/396)
Wordsmith	48.23% (4423/9170)	48.50% (796/1773)	30.35% (448/1476)	48.06% (869/1808)
Condition 2: Skipped true-positive tweets				
Dataset	PER	ORG	LOC	MISC
Finin	33.00% (435/1318)	34.83% (527/1513)	31.99% (381/1191)	-
Ritter	34.12% (1528/4478)	44.00% (898/2041)	37.11% (1305/3517)	50.67% (2067/4079)
MSM 2013	23.57% (1633/6928)	28.09% (545/1940)	30.67% (196/639)	35.99% (203/564)
Wordsmith	50.86% (2952/5804)	44.83% (473/1055)	35.22% (329/934)	50.05% (514/1027)

Table 6.8: *Skipped Tweets* - Skipped tweets containing entities

Skipped Tweets: Sentiment Analysis

Table 6.9 summarises the sentiment distribution of positive, negative and neutral tweets in the different datasets. The results present the Finin, Ritter and MSM corpora as having slightly more positive than negative tweets. The Wordsmith corpus had more tweets with negative sentiments than positive. It is worth noting here that the tweets marked negative did not necessarily have to be an aggressive or abusive tweet. An example of a tweet with a negative sentiment from the Ritter dataset is '*It's the view from where I'm living for two weeks. Empire State Building = ESB. Pretty bad storm here last evening*'. The next set of results in Table 6.10 highlights the relationship between skipped tweets and their content sentiment. The result reveals marginally that tweets with a positive sentiment were more likely to be skipped. This is inconclusive as it does not evidence to a highly polarised set as a result of the sentiment distributions.

Annotation Time: On Correct Annotations

Table 6.11 contains the average time taken for a worker to correctly identify a single occurrence of the different entity types. The results for the Finin, Ritter and MSM2013 datasets consistently present the shortest time needed corresponds to annotating locations, followed by person entities. In the Wordsmith dataset, workers correctly identified people instances in the shortest time overall, however, much longer times were taken to

Sentiment Analysis				
Dataset	POS	NEG	NEU	UNK
Finin	41.04% (181/441)	38.10% (168/441)	20.63% (91/441)	00.23% (1/441)
Ritter	47.12% (1128/2394)	36.05% (863/2394)	15.96% (382/2394)	00.88% (21/2394)
MSM 2013	40.14% (582/1450)	34.48% (500/1450)	24.62% (357/1450)	00.76% (11/1450)
Wordsmith	36.69% (1240/3380)	46.45% (1570/3380)	16.01% (541/3380)	00.85% (29/3380)

Table 6.9: *Sentiment Analysis* - General distribution

Condition 1: Sentiment Analysis				
Dataset	POS	NEG	NEU	UNK
Finin	39.75% (64/161)	36.65% (59/161)	20.63% (38/161)	(0/161)
Ritter	38.28% (694/1813)	46.83% (849/1813)	14.62% (265/1813)	(5/1813)
MSM 2013	43.00% (562/1307)	28.84% (377/1307)	27.16% (355/1307)	(13/1307)
Wordsmith	41.98% (1508/3592)	41.25% (1482/3592)	16.31% (586/3592)	(16/3592)
Condition 2: Sentiment Analysis				
Dataset	POS	NEG	NEU	UNK
Finin	45.89% (407/888)	33.03% (293/888)	21.08% (187/888)	(1/888)
Ritter	49.67% (1895/3815)	31.66% (1208/3815)	18.03% (688/3815)	(24/3815)
MSM 2013	42.16% (729/1729)	31.52% (545/1729)	25.45% (440/1729)	(15/1729)
Wordsmith	43.25% (1150/2659)	37.57% (999/2659)	18.65% (496/2659)	(14/2659)

Table 6.10: *Skipped Tweets* - Sentiment analysis distribution of skipped tweets

identify places. This result was consistent across the 2 experiment conditions with workers consistently taking shorter times to identify location and person entities. The results however note that workers took shorter time in identifying all entity types in C2 as compared to C1. Workers took on average 1 second less to identify entities in C2. In both experiment conditions, the miscellaneous entity type took the longest time to be identified taking almost 2 seconds longer on the average as compared to location entities. We posit that the extended annotator guidelines contributed to the decrease in annotation time. As this was the variable in this condition, our hypothesis is that a more detailed level of annotation guidelines leads to an anchored and increased confidence amongst the annotators. This in turn leads to mechanistic annotations – i.e. spotting a text and annotating it according to the guideline without discerning the relevant context. This can explain for the increase in speed which did not necessarily result in an increase in annotation quality.

Condition 1: Avg. Annotation Time				
Dataset	PER	ORG	LOC	MISC
Finin	9.54	12.15	8.91	-
Ritter	9.69	10.05	9.35	10.88
MSM	9.54	10.77	8.70	10.35
Wordsmith	8.06	8.50	9.56	9.48
Condition 2: Avg. Annotation Time				
Dataset	PER	ORG	LOC	MISC
Finin	7.20	7.05	6.94	-
Ritter	8.70	9.01	8.65	10.22
MSM	7.73	8.75	7.76	9.69
Wordsmith	6.88	6.79	6.97	8.72

Table 6.11: *Experiment results* - Average accurate annotation time

Interface and Heatmaps

Figure 6.6 visualises the result of our datapoint captures via heatmaps. The results presents mouse movements concentrated horizontally along the length of the tweet text area. Much activity is also around the screen center where the entity text appears after it is clicked. The heatmaps then diverge in the lower parts of the screen which indicate which entity types were tagged. From a larger image of the interface in Figure 6.2, we can reconcile the mouse movements to point predominantly to PER and LOC entities in proportions which are consistent with the individual numbers presented in Table 6.4.

Average Position of First Entity		
Dataset	Gold Entity	User Entity
Finin	16.91	22.93
Ritter	34.56	22.81
MSM 2013	35.61	24.77
Wordsmith	14.68	21.33

Table 6.12: *Experiment results* - Average Position of First Entity

A corollary to the visualisation presented in the heatmaps is the result outlined in Table 6.12. The results contain the average position of the first entity in the dataset gold standard and the average position of the first entity annotated by the workers. From the results we note that although the average positions in the gold standards vary from the 14th character in the Wordsmith dataset to the 35th character in the MSM dataset, the average worker consistently tagged the first entity around the 21st to 24th character mark. This result was consistent across all the four dataset and in variance with the results from the gold standards. We would shed more light into this in the discussion section.

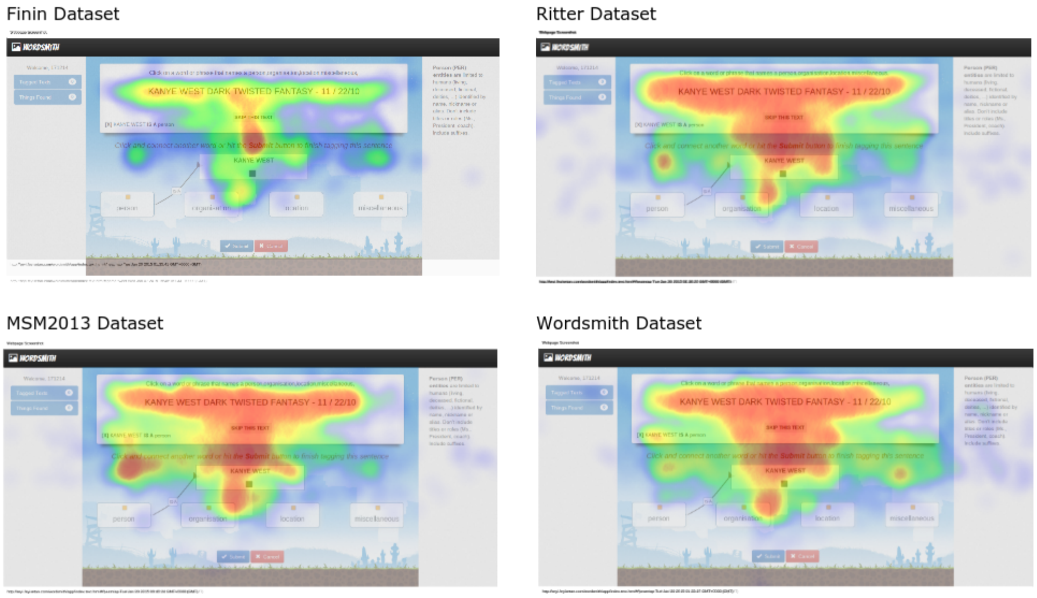


Figure 6.6: Wordsmith Heatmaps across the 4 datasets

Inter-Annotator Agreement

Table 6.13 summarises the average inter-annotator agreement scores across the four datasets. Based on our design choices, workers were allowed to skip tweets which they could not confidently annotate. Workers were required to annotate at least 10 tweets and each tweet was annotated by at least 3 annotators. The results presented here represent the inter-annotator agreement on tweets which were annotated by 3, 4, 5 and 6 workers each. At a high level, the results suggest that agreement begins to break down as consensus is required amongst more workers. This is not surprising as a base agreement between 2 out of 3 workers is equivalent to 66.67%. Drawing workers out of the same distribution on a tweet annotated by 4 workers yields a lower score of 50%. This interprets the decline in inter-annotator agreement scores as more workers annotated the same tweet.

Dataset	Number of Annotators			
	3	4	5	6
Finin	**62.40**	53.84	48.39	49.47
Ritter	**62.28**	52.84	47.11	39.03
MSM	**83.47**	83.08	79.80	77.86
Wordsmith	**60.28**	57.03	50.16	41.90

Table 6.13: *Experiment results* - Average Inter-Annotator Agreement

The inter-annotator agreement scores were clearly highest in the MSM 2013 dataset (83.47%). This can be attributed to the relative homogeneity of the dataset and the presence of a large number of easily identifiable PER entities. The other 3 datasets had similar scores with an average inter-annotator agreement of 61.65% and a standard deviation of 1.19.

Entity Inter-Annotator Agreement				
Dataset	PER	ORG	LOC	MISC
Finin	51.68	23.07	47.95	18.27
Ritter	68.05	13.67	34.14	14.69
MSM	86.95	13.20	33.72	10.62
Wordsmith	70.68	13.47	40.38	11.42

Table 6.14: *Experiment results* - Entity Level Inter-Annotator Agreement

In Table 6.14, we drill further into the inter-annotator agreement on the entity level. The results presented in this table were based on the results of 3 annotators per tweet (extrapolated from the first column in the results within Table 6.13). The results are in line with earlier results presented i.e. workers are better at identifying PER and LOC entities (as these entity types receive the highest scores), and have greater difficulties with ORG and MISC entities.

An agreement threshold of 2 workers was beneficial for the precision of identifying all the entity types across all datasets. This effect was strongest in the Wordsmith dataset where a minimum threshold of 2 raised the precision scores of identifying organisations by 20%. The least significance of the inter-annotator threshold was in identifying miscellaneous entity types in the MSM dataset where the precision score moved up by barely 0.5%. The recall values for identifying locations were the most enhanced by setting a threshold agreement of at least 2 workers. The raise in recall also signalled the least gain in the miscellaneous entity types in the MSM dataset.

Increasing the agreement threshold to at least 3 workers marked a further surge consistent with the results from setting a threshold of 2. The highest precision scores are also from the Wordsmith dataset in identifying organisations which had a boost of 30%. Precision scores in the MSM and Ritter datasets also went up over 20% by setting the inter-annotator worker threshold to a minimum of 3. As with the results presented in the previous paragraph, the lowest precision and recall score enhancements came from annotating miscellaneous entity types in the MSM dataset.

6.6.2 Summary of Findings

6.6.2.1 Overview

The low performance values for the Ritter dataset can be attributed in part to the annotation schema – just as in Derczynski et al. (2015). For example, the Ritter gold corpus assigns the same entity type *musicartist* to single musicians and group bands.

Finin dataset				
	Inter Annotator ≥ 2		Inter Annotator ≥ 3	
Entity	Precision	Recall	Precision	Recall
PER	2.77	4.69	2.12	4.61
ORG	7.65	3.33	9.17	5.37
LOC	**8.74**	**9.17**	**12.45**	**13.01**
MISC	-	-	-	-
Ritter dataset				
	Inter Annotator ≥ 2		Inter Annotator ≥ 3	
Entity	Precision	Recall	Precision	Recall
PER	5.11	5.17	9.83	7.65
ORG	**14.60**	4.62	**22.85**	5.74
LOC	11.58	**6.92**	16.46	**10.52**
MISC	14.35	3.79	22.37	2.62
MSM2013 dataset				
	Inter Annotator ≥ 2		Inter Annotator ≥ 3	
Entity	Precision	Recall	Precision	Recall
PER	5.38	4.53	6.37	6.10
ORG	**15.33**	3.66	**21.18**	4.12
LOC	11.67	**8.52**	14.72	**9.99**
MISC	0.49	1.12	0.60	-3.34
Wordsmith dataset				
	Inter Annotator ≥ 2		Inter Annotator ≥ 3	
Entity	Precision	Recall	Precision	Recall
PER	11.30	**9.09**	14.16	**13.76**
ORG	**20.49**	2.34	**29.69**	0.77
LOC	10.15	7.07	13.28	10.06
MISC	10.68	2.64	31.97	0.56

TABLE 6.15: *Inter Annotator Deltas* - Change in precision and recall values based on different inter-annotator thresholds

More significantly, the dataset does not annotate Twitter *@usernames* and *#hashtags*. Considering that most *@usernames* identify people and organisations, and the corpus contained 0.55 *@usernames* per tweet (as listed in Table 6.1), it is not surprising that scores are rather low. The result also reveals high precision and low confusion in annotating locations, while the greatest ambiguities come from annotating miscellaneous entities.

The Finin dataset has higher F1 scores across the board when compared to the Ritter experiments. The dataset did not consider any MISC annotations and although it includes *@usernames* and *@hashtags*, only the *@usernames* are annotated. Here again, the best scores were in the identification of people and places.

For the MSM2013 dataset highest precision and recall scores were achieved in identifying PER entities. However, it is important to note that this dataset (as highlighted in Table 6.1) contained, on average, the shortest tweets (88 characters). In addition, the URLs, *@usernames* and *#hastags* were anonymized as _URL_, _MENTION_ and _HASHTAG_, hence the ambiguity arising from manually annotating those types was removed. Furthermore, the corpus had a disproportionately high number of PER entities (1,126 vs. just 100 locations). It also consisted largely of clean, clearly described, properly capitalised tweets, which could have contributed to the precision. Consistent with the results above, the highest scores were in identifying PER and LOC entities, while the lowest one was for those entities classified as miscellaneous.

Our own *Wordsmith dataset* achieved the highest precision and recall values in identifying people and places. Again, crowd workers had trouble classifying entities as MISC and significant noise hindered the annotation of ORG instances. A number of ORG entities were misidentified as PER and an equally high number of MISC examples were wrongly identified as ORG. The Wordsmith dataset consisted of a high number of *@usernames* (0.55 per tweet) and the highest concentration of *#hashtags* (0.28 per tweet).

Disambiguating between ORG and LOC types remained challenging across all datasets as evidenced in the confusion matrices in Table 6.5. Identifying locations such as *London* was a trivial task for contributors, however, entities such as museums, shopping malls, and restaurants were alternately annotated as either LOC or ORG. Disambiguating tech organisations was not trivial either – that is, distinguishing entities such as Facebook, Instagram, or Youtube as Web applications or independent companies without much context. In the Wordsmith dataset, however, PER, ORG, and MISC entity tweets were skipped with equal likelihood. This is likely due to a high number of these entities arising from *@usernames* and *#hashtags*, as opposed to well-formed names. As noted earlier, this was a characteristic of this dataset, which was not present in the other three.

6.6.2.2 Analysis of tweet features

We now discuss our results in light of H1 RQ1.1 which states that specific features of microposts affect the accuracy and speed of crowdsourced entity annotation. We present these results in light of tweets which were annotated correctly, incorrectly and skipped tweets. We focus on four main features:

1. the number of entities in the micropost;
2. the type of entities in the microposts;
3. the length of micropost text;
4. the micropost sentiment

Number of entities

From the results in Table 6.7 we see that the number of entities in a tweet affect the likelihood of annotation by a worker i.e., regardless of whether the annotations are accurate or not, a tweet with fewer entities was more likely to be selected. We note that workers were more likely to annotate tweets which had fewer entities than the dataset average as contained in Table 6.1. This is further seen in the lower recall scores (as compared to precision) in Table 6.4; workers are more likely to annotate one entity in a tweet, or completely ignore tweets which have more entities than the dataset average. Workers therefore skipped longer tweets more frequently.

The results in Table 6.6 give further insight into the role of the number of entities in correctly and incorrectly annotated tweets. The results points out consistently across the 4 datasets that once a tweet has been selected for annotation, it is more likely to be annotated correctly and completely if it has fewer entities, while tweets with more entities were wrongly annotated. In summary, skipped tweets (more entities), incorrect tweets (less than skipped tweets), correct tweets (even less than both).

Types of entities

Table 6.8 and Figure 6.5 give details on skipped true positive tweets and the corresponding entity distributions. The table indicates for each dataset the total entity type encounters by the crowd workers and how many were skipped. For the first experiment condition C1 with the baseline annotation guidelines, workers skipped tweets that contained ORG entities with the highest frequency. Comparing this with our dataset overview in Table 6.1, we observe that even though the ORG type was not the most common entity type in any of the datasets, yet it was the most skipped. The next most skipped entity type was the MISC entity type in the MSM and Ritter corpora (there were no MISC annotations in the Finin gold standard). The Wordsmith dataset had the PER, ORG and MISC entity types skipped with equal frequency. For the Wordsmith dataset, as discussed earlier, this can be attributed also to entities arising from *@usernames* and *#hashtags*. The other datasets either exclude them or do not annotate them in their gold standards.

In the second experiment condition C2, in which workers were given further instructions on how to disambiguate entity types such as restaurants and museums as LOC; and universities, sport teams and musical bands as ORG, workers were then less likely to skip this entity type. Even though this did not raise precision and recall scores (as seen in Table 6.4), workers did not skip the ORG entity types as often as they did without the instructions. 3 of the 7 extra instructions explained in some form how to identify ORG entities and this likely contributed to them being skipped less. In C2, the MISC entity type was the most skipped on the average. People-related tweets were skipped more in the Finin and Wordsmith dataset, but this is a function of the high number of entities of this type (see also Table 6.1) rather than an indicator of crowd behaviour. The MSM dataset had a high number of PER entities, however, these were not skipped as the tweets were from well structured texts e.g., quotes with the author attribution at the end.

Micropost text length

The resuslts presented in Table 6.6 and Table 6.7 suggest that the tweet length was a factor in determining whether it was selected for annotation or not (since workers were free to select what tweet they annotated). However, after the tweet has been selected, there was no strong connection between the length of the tweet and the annotation accuracy. The standard deviation of the datasets was 5.65 characters, however, the standard deviation of tweets selected for annotation was 3.41 characters. As a result, at the selection stage, the tweet length played a role in the likelihood of a worker deciding to annotate, however, the length did not further matter as most of the tweets were of similar lengths.

Table 6.7 reveals that workers prefer tweets with fewer characters. The Ritter dataset with a mean tweet length of 102 characters had workers annotating posts which hovered slightly above this average length. The MSM2013 dataset had the shortest tweets with an average length of 88 characters, however, workers were willing to annotate annotate tweets with up to 9 characters above the corpus average. The Finin and Wordsmith datasets both had tweets with an average length of ≈98 characters with workers annotating similarly around this average point.

These results are reinforced in C2 with workers annotating tweets in the 98-99 character length set and discarding tweets over 100 characters. This result was consistent in all datasets asides the Ritter dataset, which had an overall set of longer tweets. From this we observe that regardless of the dataset (such as the MSM dataset with an average length of 88 characters), workers would be willing to annotate up to a certain threshold before they start skipping.

These results might not be unconnected with the user interface design. Revisiting our interface in Figure 6.3 gives an insight into how the tweets appear in the annotation interface. Shorter tweets would fit squarely in the task box with minimal text wrapping. This layout is similar to Bontcheva et al. (2014b) in that the GATE annotation tool also lays out the tweet horizontally (for workers to annotate from left to right) unlike Finin et al. (2010) which lays the tweet vertically (for workers to annotate from top to bottom). Interpreting this further in the light of the results in Table 6.12 might suggest that workers were annotating entities immediately within their field of vision since they consistently started annotating at a given point across all the datasets.

Micropost sentiment

Our experiments indicate marginally that tweets with a positive sentiment were more likely to be skipped. This is inconclusive, as it does not illustrate a polarised set as a result of the sentiment distributions. It might be possible to study the effect of tweet sentiment in annotations by carrying out granular sentiment analysis, categorising tweets as nervous, tense, excited, depressed, rather than assigning the generic positive, negative and neutral labels. Sentiment features might also be prominent in a dataset that features deleted tweets, flagged tweets or reported tweets. Other potential classes might be tweets posted to celebrities or tweets during sporting events and concerts.

6.6.2.3 Analysis of behavioural features of crowd workers

We now discuss our results in light of H2 RQ2.1, which states that we can understand crowd workers preferences based on:

1. the number of skipped tweets (with entities that could have been annotated);
2. the precision of answers;
3. the amount of time spent to complete the task;
4. the worker interface interaction

Number of skipped tweets

Tables 6.7, 6.8, and 6.10 give insights into the skipped tweets. The results indicate that across the datasets, the number of entities and the length of the tweet were two factors that contributed to the likelihood of a skipped tweet. Table 6.8 further highlights the role entity types play on workers choosing to annotate a tweet or not. At this time we cannot present conclusive remarks on the effect of the tweet sentiment on a workers probability of annotating it.

Apart from these high level features such as the number and type of entities, and the micropost length, we also discovered some other latent features which might contribute to workers skipping tweet. For example, a closer look at the Wordsmith dataset (which was the most recent corpus) revealed that workers skipped the various entity types with almost equal likelihoods. We reported this as being tied to an increase in the use of *#hashtags* and *@mentions*. Furthermore, the corpus contained #hashtags referencing events such as the #WorldCup2014 and #LondonFashionWeek which created annotation ambiguity. In the second experiment condition C2, workers spent less time annotating and skipped fewer entities due to the availability of detailed guidelines. As noted earlier, this helped workers disambiguate some entity types (e.g. handling entities from #hashtags), however, it did not result in an overall improvement in annotation quality.

Accuracy of answers

From the results in Table 6.4 we note that the crowd workers were better at identifying PER and LOC entities, and poor at characterizing MISC entity types. Table 6.5

gives further insights into the mismatching between organisation and locations (e.g., restaurants), organisations and persons (e.g., musical bands) and organisations and miscellaneous entities.

Amount of time spent to complete the task

As listed in Table 6.11 locations and people are quickly identified. In addition, the tagging speed goes up with an expansion in annotation guidelines (although the accuracy remains constant or even declines slightly). Tweets with MISC entities took the longest time to be annotated.

Worker interface interaction

We presented the findings from our heatmap datapoints in the result section and visualised them in Figure 6.6. Table 6.12 further implies to us that, workers tend to start annotating around a specific start point. In our experiments, we discovered that regardless of the dataset, workers started labelling entities that occurred around the 21st to 24th character. The Finin and Wordsmith dataset however had much lower start points in their gold standard (after 15 characters) while the Ritter and MSM corpora had much higher ones (after 35 characters). We took into consideration the responsive nature of the interface which could have presented the annotation text slightly different on varying screen resolutions and with screen resizing, and ensured that the micropost texts were presented in the same way on various screens.

Implicitly named entities

In our investigation we paid special attention to those entities that were annotated by the crowd but that were not covered by the gold standard. As a result of a manual inspection of these cases one particular category of entities stands out, which we call *implicitly named entities*. By that term we mean those entities that were represented in the text by a proxy phrase that – if the user's contextual assumptions are known – one can infer an actual named entity. A particular example for this is the annotated phrase *'last stop'*, which, if one would know the place, direction and means of transportation to contextualize the annotation, could be resolved to one explicit stop or station.

6.7 Discussion

In this final section we assimilate our results into a number of key themes and discuss their implications on the prospect of hybrid NER approaches that combine automatic tools with human and crowd computing.

Crowds can identify people and places, but more expertise is needed to classify other entities

Our analysis clearly reveals that microtask workers are best at spotting locations, followed by people, and finally with a slightly larger gap, organisations. When no clear instructions are given, that is, when the entity should be classified as MISC, the accuracy suffers dramatically. Assigning entities as organisations seems to be cognitively more complex than persons and places, probably because it involves disambiguating their purpose in context e.g., universities, restaurants, museums and shopping malls. Many of these entities could also be ambiguously interpreted as products, brands, or even locations, which also raises the question of more refined models to capture diverse viewpoints in annotation gold standards Aroyo and Welty (2013). To improve the crowd performance, one could imagine interfaces and instructions that are bespoke for this type of entities. However, this would assume the requester has some knowledge about the composition of his corpus and can identify problematic cases. A similar debate has been going on in the context of GWAPs, as designers are very restricted in assigning questions to difficulty levels without pre-processing them Simperl et al. (2013). One option would be to try out a multi-step workflow (such as the hybrid workflow proposed by Sabou et al. (2013)) in which entity types that are empirically straightforward to annotate are solved by 'regular' workers, while miscellaneous and other problematic cases are only flagged and treated differently – be that by more experienced annotators, via a higher number of judgements Snow et al. (2008), or otherwise.

Crowds perform best on recent data, but remember people

All four analysed datasets stem from different time periods (Ritter from 2008, Finin from 2010, MSM from 2013, and Wordsmith from 2014). Most significantly one can see that there is a consistent build-up of the F1 score the more recent the dataset is, even if the difference is only a couple of months as between the MSM and the Wordsmith cases. We interpret that the more timely the data, the better the performance of crowd workers, possibly due to the fact that newer datasets are more likely to refer to entities that gained public visibility in media and on social networks in recent times and that people remember and recognize easily. This concept known as entity drift was also highlighted by Derczynski et al. (2015) and Fromreide et al. (2014). The only exception for this is the PER entity type, which was the most accurate result for the MSM dataset. However, in order to truly understand this phenomenon we would need more extended experiments, focusing particularly on people entities, grounded in cognitive psychology and media studies (Cheng et al., 2013; Minkov et al., 2005).

Partial annotations and annotation overlap

The experiments hint at a high share of partial annotations by the workers. For example, workers annotated *london fashion week* as *london* and *zune hd* as *zune*. Other partial annotations stemmed from identifying a person's full name, e.g., *Antoine De Saint Exupery* was tagged by all three annotators as *Antoine De Saint*. Overlapping entities occurred when a text could refer to multiple nested entities e.g., *berlin university museum* referring to the university and the museum and *LPGA HealthSouth Inaugural Golf Tournament* which was identified as an organisation and an event. These findings call for richer gold standards, but also for more advanced means to assess the quality of crowd results to reward partial answers. Such phenomena could also signal the need for more sophisticated microtask workflows, possibly highlighting partially recognized entities to acquire new knowledge in a more targeted fashion, or by asking the crowd in a separate experiment to choose among overlaps or partial solutions.

Spotting implicitly named entities thanks to human reasoning

Our analysis revealed a notable number of entities that were not in the gold standard, but were picked up by the crowd. A manual inspection of these entities in combination with some basic text mining has shown that the largest set of these entities suggest that human users tend to spot unnamed entities (e.g., *prison* or *car*), partial entities (e.g., *apollo* versus *the apollo*), overlapping entities (e.g., *london fashion week* versus *london*), and hashtags (e.g., *#WorldCup2014*). However, the most interesting case were the ones we call *implicitly named entities*. Examples such as *hair salon*, *last stop*, *in store*, or *bus stop* give evidence that the crowd is good at spotting phrases that refer to real named entities implicitly depending on the context of the post's author or a person or event this one refers to. In many cases, the implicit entities found are contextualised within the micropost message, e.g., *I'll get off at the stop after Waterloo*. This opens up interesting directions for future analysis that focus only on those implicit entities together with features describing their context in order to infer the actual named entity in a human-machine way. By combining text mining and content analysis techniques, it may be possible to derive new meaning from corpora such as those used within this study.

Closing the entity recognition loop for the non-famous

Crowd workers have demonstrated good performance in annotating entities that were left out by the gold standards and presented four characteristic classes of such entities: (i) unnamed entities, (ii) partial entities, (iii) overlapping entities, and (iv) hashtags. It is noteworthy that we observed an additional fifth class that human participants mark as entities, which refer to non-famous, less well-known people, locations, and organisations (e.g., the name of a person who is not a celebrity or a place in a city that would not fall into the category of a typical point of interest). This is an important finding for hybrid entity extraction pipelines, which can benefit from the capability to generate new URIs for yet publicly unknown entities. This can play an important role in modern (data) journalism (Luczak-Rösch and Heese, 2009) and complements the findings about the entity annotation behaviour of technical non-experts on longer texts presented in

Hinze et al. (2012a) and Hinze et al. (2012b).

Wide search, but centred spot

Our heatmap analysis indicated that we had a very wide view along the text axis, and a consistent pattern that the likelihood of annotating in the centre is higher even though they seem to search over the entire width of the text field. This correlates with statistics about the average position of the first annotation, which remained constant in the user annotations as compared to the varying positions in the gold standard. Workers started off by annotating entities at the beginning of the tweet then around the middle of the tweet before the tagging recall dropped. This might mean that people are more likely to miss out on annotating entities on the right edges of the interface or at the end of the text. A resolution could be to centralize the textbox and make it less wide hence constraining the worker's field of vision as opposed to Finin et al. (2010) where workers were required to observe vertically to target entities.

Useful guidelines are an art

Our study seems to indicate that additional instructions do not always produce better tagging quality. We noted, however, that it has the following effects: (i) it speeds up the annotation process as we noted that workers on the average spent less time annotating entities; (ii) it makes people more willing to undertake choice-based work – tweets with ORG entities were less skipped after the introduction of more detailed guidelines. However, this did not affect the accuracy scores, which were in fact reduced in a few places. The new guidelines did not remove worker bias towards identifying implicit unnamed entities. Workers continued to tag concepts such as room, gym and on the road as entities even when the instructions tried to discourage them to do so. While giving effective feedback is an ongoing research problem in crowdsourcing, one approach, which we could investigate more is crowd-based feedback and crowd sociality, using synchronous work by workers who are completing tasks in the same time. A previous study we carried out (Feyisetan et al., 2015b) points out that crowd workers appreciate features which offer continuous feedback mechanisms and a view into how other workers are performing with the task. Another interesting question would be if we could leverage the efforts people invested in tagging things we were not looking for. While it is clear that crowdsourcing, at least on paid microtask platforms, is goal-driven and that the requester is the one setting the goals, it might make sense to consider models of co-creation and task autonomy, in which as the tasks are being completed, the requester takes into account the feedback and answers of the crowd and adjusts the goals of the project accordingly. Literature on motivation tells us that people perform best when they can decide what they are given the freedom to choose what they contribute, how, and when, and when they feel they are bringing in their best abilities (Deci and Ryan, 1985b). These aspects might not be at the core of CrowdFlower and others, which focus on extrinsic motivation and rewards, but they are nevertheless important and could make experiments more useful in several ways.

Revisiting the role of experts
Some of the results presented here might ferment questions on the usefulness of the crowd in carrying out high quality named entity recognition on noisy microposts. Indeed, the crowd is but one step in the workflow required to achieve the Web of Data vision and understanding how to harness their unique capabilities is of utmost importance. Automatic annotation processes have continued to improve and this has been in part due to the availability of pre-annotated corpora – carried out by experts and the crowd. We believe our work would form one of the missing components in addressing the design of more advanced workflows which could necessitate the reintroduction of experts into the loop – fitting in to disambiguate where the crowd falls short.

In addition, the crowd helps to shed further light into what might have been overlooked by a trained set of experts, opening up potentials out of scope of predefined research questions. For example, in our case, the potentials of implicit entities could help in the design of conversational AI assistants which could resolve *last stop*, *in store*, or *bus stop* based on context.

6.8 Introducing Furtherance Incentives

Given the insights we have garnered in the discussion section, it becomes paramount to leverage on the information to achieve our original intention of building better workflows. However, beyond building better workflows (which leads to an increase in work quality), would it be possible to also design a workflow that improves task uptake and engagement among crowd workers? Can we utilise the insights of what the crowd is good at to design a more engaging task experience?

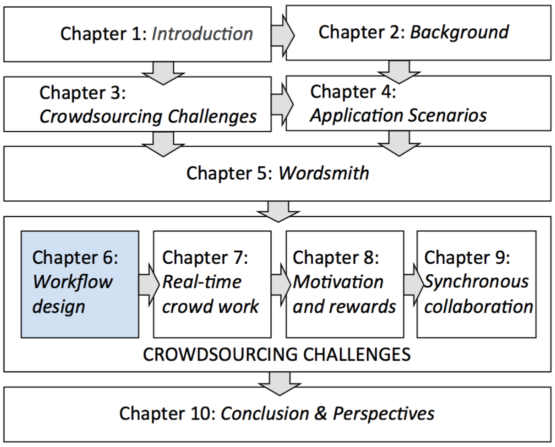
We believe the answer to this lies in a concept which we term '*furtherance incentives*'. (We provide a full description of furtherance incentives in Chapter 8). Simply explained, furtherance incentives serve as a stimulus to improve task continuance by introducing it (the incentive) at the point when a worker is about to quit a task. For example, our experiments reveal workers prefer to annotate tweets with PER entities while they perform badly on tweets with MISC entity types. In the presence of the element of choice (as was with our experiments), workers required to annotate successive tweets containing just MISC entity types would tag only the number of tweets required to receive their payment. However, workers annotating tweets with PER entities tag more tweets. Using this insight, tweets with PER entities can be used as a content based furtherance incentive for workers who are about to drop off from the task.

In the next chapter on real-time crowdsourcing, we would expand on this concept slightly – presenting how furtherance incentives can also be introduced to improve real-time crowd tasks. In Chapter 8 while addressing the challenge of motivation and rewards, we present a more rigorous definition and experimental study on improving motivation in microtasks using furtherance incentives. Finally in Chapter 9, we also apply furtherance incentives of social pressure and social flow while investigating the challenge of synchronous collaboration in microtask crowdsourcing.

6.9 Conclusion

In terms of the wider impact of our study, we consider that our findings will be useful for streamlining and improving hybrid NER workflows, offering an approach that allows corpora to be divided up between machine and human-led workforces (comprising of generic crowds, and hierarchical mediators or experts), depending on content features such as the types and number of entities to be identified or the length of the tweets. Future work in this area includes (i) devising automated approaches to determining when best to select human or machine capabilities; (ii) examining *implicitly named entities* in order to develop methods to identify and derive message-related context and meaning; as well as (iii) looking into alternative ways to engage with contributors using real-time crowdsourcing which we present in Chapter 7, crowd feedback, multi-steps workflows involving different kinds of expertise to improve tagging performance for organizations and other ambiguous entities, and giving the contributors more freedom and autonomy in the annotation process.

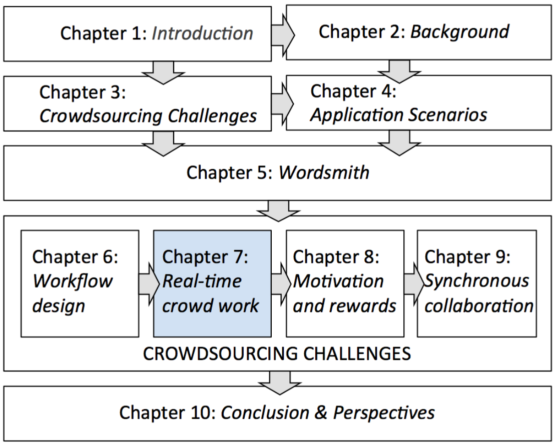
6.10 Summary


Chapter 1: *Introduction*
Chapter 2: *Background*
Chapter 3: *Crowdsourcing Challenges*
Chapter 4: *Application Scenarios*
Chapter 5: *Wordsmith*
Chapter 6: *Workflow design*
Chapter 7: *Real-time crowd work*
Chapter 8: *Motivation and rewards*
Chapter 9: *Synchronous collaboration*
CROWDSOURCING CHALLENGES
Chapter 10: *Conclusion & Perspectives*

In this chapter, we studied how understanding task content features and crowd worker abilities and preferences can be used to design better crowdsourcing workflows. We investigated an approach to finding entities within micropost datasets using crowdsourced methods. Our experiments, conducted on four different corpora, revealed a number of crowd characteristics with respect to their performance and behaviour of identifying different types of entities.

Chapter 7

Real-time Crowd Work


Chapter 1: *Introduction*
Chapter 2: *Background*
Chapter 3: *Crowdsourcing Challenges*
Chapter 4: *Application Scenarios*
Chapter 5: *Wordsmith*
Chapter 6: *Workflow design*
Chapter 7: *Real-time crowd work*
Chapter 8: *Motivation and rewards*
Chapter 9: *Synchronous collaboration*
CROWDSOURCING CHALLENGES
Chapter 10: *Conclusion & Perspectives*

In this chapter we use crowdsourcing contests which, in combination with individual micro-payments, allow us to collect judgements effectively under tight time constraints. We present our crowdsourcing contest model followed by our approach at predicting worker drop-offs. We detail our experiment setups across different reward spreads and task thresholds before highlighting our findings. Following from the previous chapter, we continue our discourse on furtherance incentives before concluding the chapter.

7.1 Overview

We extended *Wordsmith* with real-time features that allow multiple workers to compete against each other while their answers are compared and validated. Each experiment recruited a fixed number of workers from CrowdFlower, [1] an online paid microtask crowdsourcing marketplace, and had three specific constraints: time, task threshold and reward spread. Workers used Wordsmith to annotate tweets for a fixed period of *time*. In order to be eligible for payment, they had to complete a minimum number of tasks (referred to in this chapter as *task threshold*). They were rewarded only if they were high enough in the overall ranking; in other words, workers competed against each other, and only a share of them (the so-called *reward spread*) received a payoff at the end of the contest. Rankings were computed and updated on the fly as a function of the number of

[1] https://crowdflower.com

tasks completed and a heuristic approximation of their quality, based on previous work of ours on large-scale automatic named entity recognition for Twitter (Feyisetan et al., 2014).

When designing the experiments, our primary focus was to create a microtask crowdsourcing model which could be applied to different scenarios. Wordsmith as such has been used for several types of tasks, including image labelling (Feyisetan et al., 2015b) and named entity recognition (Feyisetan et al., 2015a), both for static and stream-like data. The choice of task in this chapter was not motivated by the need to design a new NER algorithm, like we did in Feyisetan et al. (2014) or as we presented previously in Chapter 6, but as a means to test our novel crowdsourcing model. Our main intuition was that by designing the crowdsourcing exercise as a live contest, which must be completed in a relatively short period of time, we create an environment in which results are delivered both fast and with accuracy. In addition, as we do not pay all workers upfront or merely for being available, we keep the overall costs lower. In this context, we hypothesized that the number of workers who would be rewarded and the amount of work that was necessary to be eligible for payment would have different effects on the quality and quantity of task output. In order to optimize unit costs further, we went on and studied exit patterns and attrition; this was very important since we were interested in the timely completion of the task as a whole, and not just in the top-k contributions, which is the case in most contests (See section 2.3.4).

We ran experiments with three sets of reward spreads (top worker; top 5 workers; and top 10 workers), and two task thresholds (low: annotate at least 1 tweet; and high: annotate at least 10 tweets). As datasets we used four benchmarks from the Twitter NER literature and previous work of ours (Feyisetan et al., 2015a). These datasets included gold standards, which were instrumental in computing crowd output quality. Our findings support our initial hypothesis: the model yields faster results ($2x$ as fast then a baseline approach from the literature). Increasing the reward spread led to an increase in task output, while a higher task threshold within each reward spread meant more work overall, but also a reduction in the contributions of the top contestants. Rewarding more workers also reduced the rate of worker attrition and kept more workers engaged.

Examining the results in detail gives an insight into when these patterns break down i.e., when an increase in reward spread does not imply increased output, or, when it actually leads to increased attrition. These insights would therefore help in finding the balance in comparing theoretic guarantees with empirical evidence to select appropriate reward spreads and task thresholds while scaling to task sizes comparable to real-world Twitter processing engines. The contest model also proved, apart from its potential financial compensation, to be an approach with intrinsic motivation. Workers not only completed more tasks than required, but some of them reported positively about the experiments on a community forum: for example, one post reads '*Hello everyone! lately*

I'm hooked on the multiplayer tasks, waiting for 100 *people to connect*', while another one claims: '*Hit the top* 10 *today. I will hunt this problem again*'.

To improve engagement and reduce the overall costs, we created a predictive model that estimates the probability of a worker exiting a contest at a certain point in time given their current task output and relative rank. This opens up the possibility of applying furtherance incentives (as introduced in Chapter 6) to discourage workers from leaving the competition.

7.2 Model

In this section, we introduce a high-level overview of our approach to crowdsource named entities in real-time. We present our microtask design model and strategies for undertaking crowd work. This involves the use of an external recruitment marketplace, CrowdFlower, and our bespoke competition platform, *Wordsmith*.

7.2.1 Task

The task consists of a total of n posts, $P = \{p_1, ..., p_n\}$, each containing m entities $E = \{e_1, ..., e_m\}$ to be annotated, where $m < M_i! + M_i$ and M_i is equal to the number of text tokens in post p_i. The posts arrive at a constant rate λ and each has a processing rate of μ. There are n workers in a pool to serve the task queue such that, to keep up with the requests, the ingress load (task intensity) $L = \lambda/\mu$ must be less than the number of workers n, i.e., $L < n$. Hence, tasks that are not solved are dropped of the queue as opposed to being kept indefinitely in the buffer (Bernstein et al., 2012). The tasks are solved using a first-in-first-out scheduling policy and processing scheme, and, already recruited workers are sought to carry out new tasks (as opposed to recruiting additional workers). Therefore, the requester is looking for an optimal processing rate μ, and needs to keep workers motivated to carry out as many tasks as possible.

In our experiments, we modelled the arrival rate λ based on previous work (Feyisetan et al., 2014) by using the average number of English tweets per second which *probably* has a named entity present (we used proper nouns as a signal indicator of the presence of named entities). We also modelled the processing rate λ using the results of a follow-up study published in Feyisetan et al. (2015a), which gave insight into the average completion rates of named entity annotation tasks by crowd workers (the results of this was reported previously in Chapter 6).

7.2.2 Constraints

The requester defines (i) a completion time constraint T, which depends on the number of posts n and their arrival rate λ; and (ii) a quality constraint Q, which denotes the minimum number of annotations expected from each worker to be eligible for payment. The latter is essential in hybrid tasks; for example, the task might have been pre-annotated by a machine to determine the probable number of named entities (this serves as the quality constraint Q), while the crowd workers identify those entities and type them (Feyisetan et al., 2015a).

7.2.3 Workers

There is a set of n workers, $W = \{w_1, ..., w_n\}$, each with the ability to carry out entity annotations, participating in the contest. Each worker $w \in W$ has a private skill level ς_i (also known as expertise or ability), and for each post in an annotation task, chooses to exert a level of effort $\epsilon_i \geq 0$. The skill level is drawn independently of other workers from the interval $\varsigma_i \backsim [0, 1], \forall w \in W$, according to a distribution function F with density $f(\varsigma) = dF(\varsigma) > 0$. The effort exerted is drawn from the interval $\epsilon_i \backsim [0, \epsilon]$, in which the maximum effort expendable is constrained by the running time t of the contest, which in turn is a function of posts per unit time and total number of posts. The quality q_i of each worker w_i is determined by the skill level ς_i, the effort exerted ϵ_i, and a requester variable δ_i. The requester variable δ_i is a function of the requester's review process and perception of quality, in comparison with the worker's internal tagging bias, which is markedly present in human judgement tasks. This value is constant across annotation posts $\forall p_i \in P$; therefore, two workers exerting the same effort to tag the same post would differ only on their skill, since the requester's variable is constant for that post. The quality of a submission is thus given as $q_i = \varsigma_i \epsilon_i + \delta_i$. In our experiments, the requester's variable was a measure of results in a pre-computed gold standard set (Feyisetan et al., 2015a).

Each worker w_i seeks to maximise their expected utility. This depends on the number and value of prizes, and the number of contestants and value of their efforts. A worker's utility U_i is given by $U_i = V_i - c(\epsilon_i)$ if the worker w_i wins prize V_i, or $U_i = -c(\epsilon_i)$ otherwise, where V_i is one of k prizes to be awarded by the requester, and, $c(\epsilon_i)$ is the worker's cost function, which is a strictly increasing function dependent on exerted effort where $c(0) = 0$. Each prize V_j above threshold k is positive, or zero otherwise; we did not model negative rewards (punishments), as they did not seem to have the desired effects in early experiments we carried out.

7.2.4 Requester

The requester asks the crowd to complete a series of tasks in real-time. The requester needs to determine the experiment setup: number of contestants, number and size of prizes, and contest constrains to maximise the effort exerted by all contestants $\sum_{i=1}^{n} \epsilon_i$. This is different from contests such as the Netflix Challenge (Bennett and Lanning, 2007), where the principal's objective was to elicit a single best response to a task. In our case the requester does not only desire to maximise total exerted effort, but also to maximise some utility function of output qualities $\sum_{i=1}^{n} q_i$. The requester therefore needs to maintain incentives for highly skilled workers, and, motivate low skilled workers to exert more effort while adjusting the prize spread.

7.2.5 Mechanism

The requester is able to observe the baseline quality of each worker's output (based on: the pre-computed number of entities, and the number of entities submitted by the worker. He/she is then able to use this information to construct a mixed cardinal-ordinal contest (Ghosh and Hummel, 2015) by assigning a quality score to every crowd answer – therefore, contestants are ranked not only based on their effort (number of posts annotated), but also on the quality of their output. The reward mechanism awards a prize A_j to an worker w_j within a reward spread (e.g., the worker was within the top 5 or top 10) if their effort level surpassed a pre-defined threshold (e.g., the worker had annotated a minimum of 5 or 10 posts).

7.2.6 Worker Exit

A worker would always seek to maximise their expected utility given the number and value of prizes, and the number of contestants and the value of their efforts. In our experiments, it was possible for workers to view their ranked position in real-time with respect to their closest contenders using a *k neighbours* leaderboard view as presented in the medium information policy contest strategy by Rokicki et al. (2014). A worker far outside the reward spread might inadvertently decide to exit the contest to avoid further loss of utility. The worker close to the reward spread might, however, decide to remain in the contest in the hopes of displacing a close contender. Queue theory tell us that the probability that all the workers are busy due to their fellow workers exiting is:

$$Pr = B(L, n) = \frac{L^n/n!}{\sum_{i=0}^{n} L^i/i!} \tag{7.1}$$

where $L = \lambda/\mu$ is the ingress load (task intensity based on the task arrival rate λ and the agent processing rate μ); and n is the number of agents. From our experiments (see Section 8.4), we observe the effect of varying the reward and reward spread not only on the task quality and output, but also on agent exit (given that a new agent is not recruited to replace an exiting one).

7.3 Crowdsourcing Design

In this section, we introduce our contest-based real-time crowdsourcing design approach. We first define the task and then present the task platform.

7.3.1 Task Description

The task consists of a total of n posts (tweets), $P = \{p_1, ..., p_n\}$, each containing m entities $E = \{e_1, ..., e_m\}$ to be annotated. During the contest, crowd workers are shown a list of tweets and they are to annotate as many of them as possible before the next set of tweets come in.

7.3.1.1 Entity Types

The required entity types were person (PER), organisation (ORG), location (LOC) and miscellaneous (MISC). These are the most common types of entities used in the NER literature from Finin et al. (2010) and Ritter et al. (2011). These were also the four named entity types reported previously in the named entity recognition task in Chapter 6, Section 6.4.

7.3.1.2 Annotation Guidelines

We presented the workers with basic information on what named entities are, and, some additional information on how to disambiguate between difficult entity classes e.g., organisations, which could be classed as locations (e.g., restaurants and museums), typos, abbreviations, colloquialisms, nested entities and software that references the name of the creating company (e.g., Instagram).

7.3.1.3 Dataset

We selected four datasets from existing literature, which we used to simulate a real-time influx of streaming tweets. More details are given in Section 7.5.

7.3.1.4 Gold Standard

Each dataset comes with an annotated gold standard which we used for evaluation purposes. The gold standard was also used to compute a contestant's accuracy on an annotation task.

7.3.2 Task Platform

The annotations were carried out on Wordsmith (see more details in Chapter 5) using the contest platform configuration. This setting allowed for multiple workers to connect to the system at once and carry out entity annotations simultaneously.

7.3.2.1 Input and Output

The system takes in a raw input of streaming posts and performs an initial sequence of processing on it. For our studies, we focus on filtering out non-English tweets using the language tag of the incoming tweets. We then carry out parts-of-speech tagging to recognise tweets with proper nouns. This is used to build a pseudo-quality score for each annotation (presented earlier as the requester's variable δ_i) i.e., if our POS tagger detects 2 different contiguous proper noun sets, we can expect an annotation result of at least 2 entities (although this does not hold strictly if there are no proper nouns e.g., entities might be recognised as noun phrases by some taggers).

The system outputs a processed stream of English tweets, *(tweet1, ..., tweetN)* where each tweet is represented as a tuple containing a reference ID, the tweet string and an associated requester's variable. Each tweet in the stream advances in linear discrete time at a constant rate with each time point represented as unique integer value in seconds (although, worker annotation and exit is represented in milliseconds).

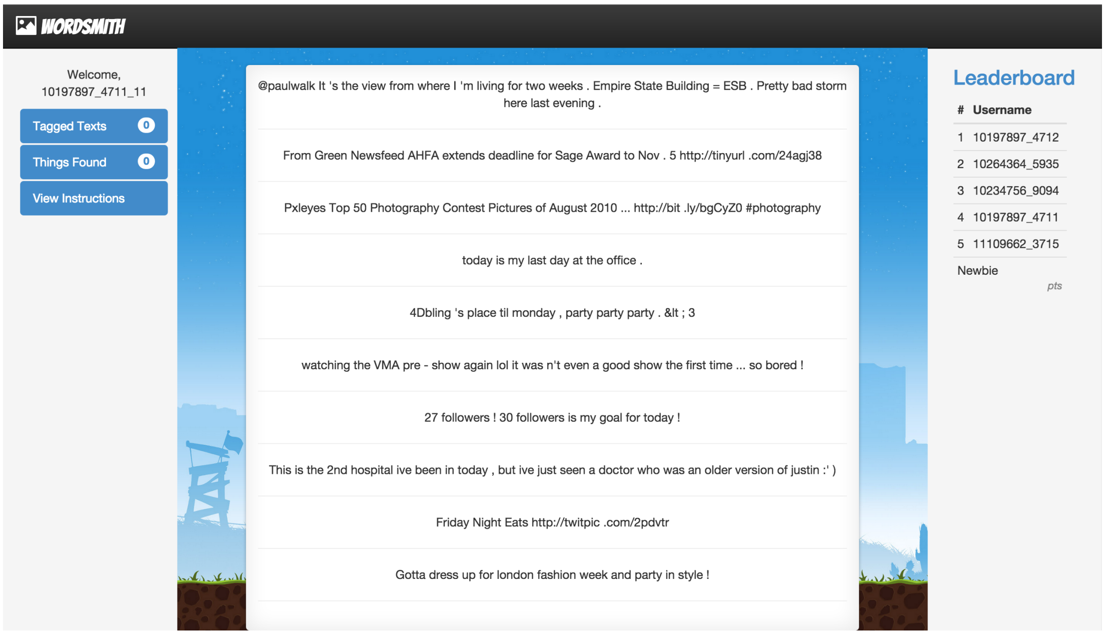
7.3.2.2 Temporal Division and Stream Parallelism

Streaming tweets are bucketed into distinct time intervals using *windows*. A window consists of a constant number of tweets which is then emitted per unit time. (which may or may not have been built over a buffer depending on varying throughput levels).

Within each window task slice, tweets are clustered and parallelized to different workers. Each cluster is a task unit consisting of a list of tweets which is allocated to each worker for annotation. This follows a Map Reduce paradigm wherein each worker has a small task unit to solve which is recursively built up to the final solution for the requester. The map process involves local processing on individual nodes (individual annotating contestants), while the reduce process involves the merging of results to select best responses for overlapping task annotations.

7.3.2.3 Interface

The task interface consisted of a central annotation panel, in which a worker saw the current list of tweets as described in Feyisetan et al. (2015a). After selecting a tweet, a worker could either mark it as having no entities, skip it or go back to the list or annotate the entities in the tweet. The worker received a baseline score x for annotating a tweet with an arbitrary number of entities and a higher score $5x$ for correctly annotating the pre-computed number of entities (based on the gold standard tags). These figures were drawn from a series of observations and preliminary experiments, which also ruled out the use of negative scores and qualifying questions, as both led to a very sharp rate of exits.



Welcome,
10197897_4711_11

@paulwalk It 's the view from where I 'm living for two weeks . Empire State Building = ESB . Pretty bad storm here last evening .

From Green Newsfeed AHFA extends deadline for Sage Award to Nov . 5 http://tinyurl .com/24agj38

Pxleyes Top 50 Photography Contest Pictures of August 2010 ... http://bit .ly/bgCyZ0 #photography

today is my last day at the office .

4Dbling 's place til monday , party party party . &lt ; 3

watching the VMA pre - show again lol it was n't even a good show the first time ... so bored !

27 followers ! 30 followers is my goal for today !

This is the 2nd hospital ive been in today , but ive just seen a doctor who was an older version of justin :')

Friday Night Eats http://twitpic .com/2pdvtr

Gotta dress up for london fashion week and party in style !

Leaderboard

#	Username
1	10197897_4712
2	10264364_5935
3	10234756_9094
4	10197897_4711
5	11109662_3715

Newbie
pts

FIGURE 7.1: Wordsmith Interface

7.3.2.4 Configuration

On the requester side, a number of configuration settings are afforded for.

- **Number of workers (*W*):** the maximum number of workers who could connect to the platform.
- **Leaderboard view:** this sets the way a worker sees other contestants on the leaderboard: *top k*, which lists the top workers on the leaderboard; or *k neighbours*,

which indicates the k contenders above and below the worker. Our experiments adopted the *k neighbours* leaderboard view based on findings from Rokicki et al. (2014).

- **Number of tweets (P):** this represents the total number of tweets to be processed.
- **Window size (w):** the number of tweets which is sent out to all workers per time slice.
- **Task unit time (μ):** the delay time, for which a list of tweets remains available for annotation to a worker before the next set of tweets arrive.
- **Task arrival rate (λ):** the number of tweets which were channelled to the platform per unit time computed as w/μ.
- **Task unit size:** the number of tweets that a worker actually sees on screen at any given time, which is a fixed percentage of the window size.
- **Total task time (T):** represents how long the contest would take. This is computed as:

  $$T = \frac{P\mu}{w} \tag{7.2}$$

  For example, in our experiments, the number of agents was 100; the number of tweets desired to be processed was $7,600$; the window size was 200; the task unit time was 10 seconds (more details under *time warping*), while the task unit size was 10 i.e., a worker was shown 10 tweets out of the current stream of 200 tweets for a period of 10 seconds (during which the worker is to annotate as many as possible before the next set of tweets arrive). The task arrival rate was $200/10$ which is 20 tweets per second and the total contest time was $(7,600 * 10)/200 = 380$ seconds (6 minutes 20 seconds).

7.3.2.5 Warping Time

Warping time is a strategy, in which a worker's task slice in a real-time assignment is deliberately slowed down to afford for maximal worker cognition in undertaking the required task. For example, Lasecki et al. (2013a) used time warping to slow down audio playback so crowd workers could effectively transcribe a given portion of speech. This is illustrated in Figure 7.2 below. This was recursively done for each worker, after which the individual results were successively merged to create a single result. Following our stream parallelism approach to dividing up the incoming microposts, a window of 200 tweets was presented to 20 workers. We adopted the approach by Lasecki et al. (2013a) denoting an in-period P_i where the annotation stream for a worker group comes in, a speed reduction rate r and the compensating out-period P_o where the worker group N

rejoins the live real-time stream. We used a speed reduction rate r of 10 i.e., the 20 workers experienced a streaming rate of $1/r = 0.10$ (i.e., one tenth speed, 10 seconds rather than 1 second) during that annotation period. During these 10 seconds, the workers previewed a static list of tweets, from which they can select individual entities to annotate. The stream of tweets in the buffered out period were then emitted at a speed of:

$$\frac{N-1}{N-r} \tag{7.3}$$

In our experiments we made certain assumptions, which would be handled differently on the live data feed. For example, all members of a worker group (20 workers in our case) were presented with the annotation tasks at the same input period P_o, which was a function of our streaming methodology. We re-purposed existing datasets from literature into a streaming API. In actual practice, each worker would have a unique input period P_o similar to Lasecki et al. (2013a).

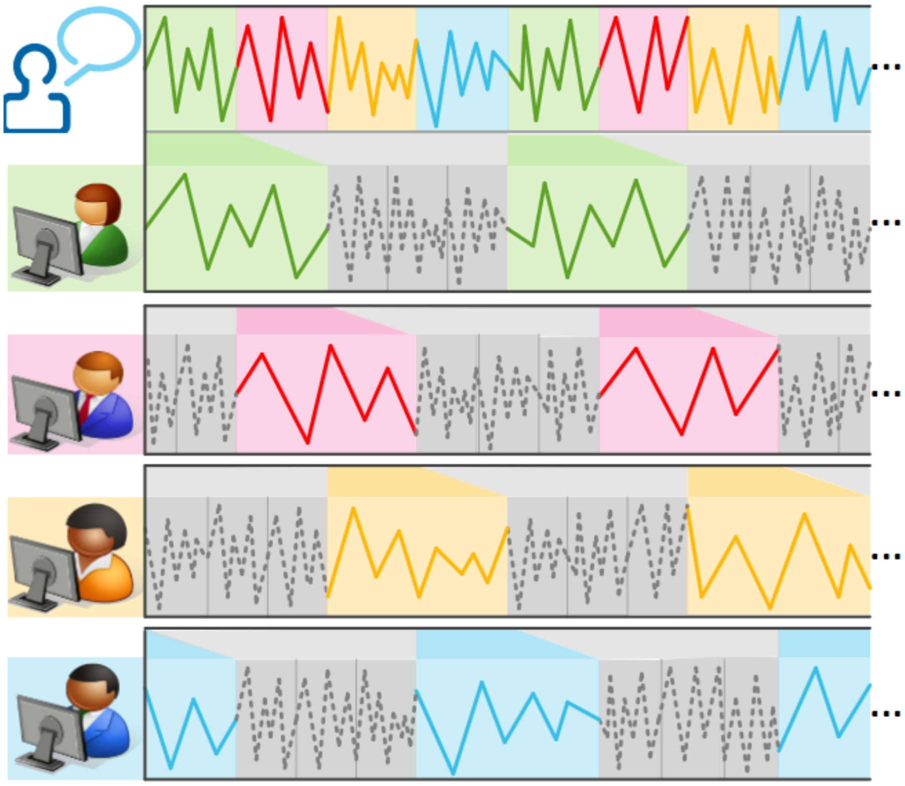


FIGURE 7.2: Warping time in a transcription task by Lasecki et al. (2013a)

7.3.2.6 Task Allocation

The problem of task allocation or task routing constitutes a research topic in its own right. There are several approaches to distributing the incoming stream of tasks to the available workers in parallel. In a *random* assignment strategy, every available contestant is assigned a random task unit slice from the window of current tweets, e.g., in our experiments, a random strategy would assign each worker 10 tweets from 1 out of 20 bins derived from the streaming window of 200 tweets. In a *round robin* assignment strategy, each bin would be sequentially assigned to the next available worker. Other task allocation strategies attempt to optimise the output by assigning tasks based on worker skill and task difficulty, by routing tasks to potentially obtain the highest information gain (Bragg et al., 2014), or by implementing it as a Markov decision process (Kobren et al., 2015). In our experiments, we adopted a random task allocation strategy.

7.4 Predicting Contest Exit

We adopt a Bayesian probabilistic reasoning approach to determine if a worker would exit the contest given the time spent and the worker's expected utility. Using this, we can create a model, which we use to predict the probability of (a number of) workers exiting at various reward spreads, at different task thresholds, and at various times in the contest. This model is computed from:

```
Result: Contest exit: $\arg\max_x \Pr(x|U)$
Parameter: U = utility;
Task threshold: $v$ = {1, 10};
Reward spread: $s$ = {1, 5, 10};
for time $t > 0$ do
    Count entity annotations $\epsilon$ as $f(A)$;
    Compute quality $q \; \forall \; a_i \in A \Rightarrow f(a_i, \delta_i, \varsigma_i)$;
    Compute cost $c = f(\epsilon, q, v)$;
    Update position $p = f(r, s)$ where $r = f(\epsilon, q)$;
    Worker Utility $U = f(c, p, t)$;
    if $U > 0$ and $U \in s$ then
        return 0;
    else
        return $\Pr(U|x) \Pr(x)$ at $t$;
    end
end
```

Algorithm 1: Contest Exit at utility U

- **The prior probability** of workers exiting the contest at various time points. This was collected empirically from the exit distributions from our experiments and presented in Figure 7.12.
- **The likelihood probability** of workers exiting the contest given their current expected utility at a given time t. This is built from a joint probability of the various parameters which comprise the worker's utility at any given time, details of which are presented below.

Beginning with simple Bayesian reasoning, we have the posterior proportional to the likelihood and the prior, where we have earlier stated that, the likelihood is a joint probability of the utility variables:

$$\Pr(x|U) = \frac{\Pr(U|x)\Pr(x)}{Pr(U)} \propto \Pr(U|x)\Pr(x)$$

given the worker's expected utility $U = f(c, p, t)$ and
the worker's cost $c = f(\epsilon, q, v)$
ϵ represents the worker's effort in terms of annotation counts
q represents the task quality score in terms of correct annotations
v represents the task threshold i.e., min annotations required
the worker's position $p = f(r, s)$
r represents the worker's rank based on ϵ and q
s represents the reward spread, i.e., the total workers to be paid
the elapsed time $t > 0$

$$\Pr(x|t, c, p) = \frac{\Pr(p, c, t|x)\Pr(x)}{\Pr(p, c, t)} \propto \Pr(c, p, t|x)\Pr(x) \tag{7.4}$$

where:
$\Pr(x|t, c, p)$ is the *posterior* of the worker's exiting the contest, given the time spent, the worker's expended cost and current position
$\Pr(x)$ is the *prior* probability of worker's exiting the contest at this particular time
$\Pr(p, c, t|x)$ is the *likelihood* at the current time that the worker would exit the contest.

Predicting the probability for a workers exit then is:

$$\Pr(x) = \int_u \Pr(x|U)P(U)dU \tag{7.5}$$

For each variable in the joint probability $\Pr(U) = \Pr(p, c, t)$, computing the likelihood probability conditioned on a worker's exit is calculated empirically at each time point by comparing the parameter value at that time, with all the values observed over the entire contest period. For example, the likelihood at time t, that a worker that has incurred cost c would exit the contest, is the integral (over all worker participation) of a unit cost observation at that time t divided by the sum of all cost incurred for the contest span, simplified as:

$$\Pr(c|x, t) = \frac{\Pr(c|x, t)}{\sum_{t=0}^{T}\Pr(c|x, t)} \tag{7.6}$$

7.5 Experiment Design

We used CrowdFlower to source and remunerate workers crowd workers. Each CrowdFlower job included a link to Wordsmith.

7.5.1 Research Questions

We sought to answer four research questions:

1. Can the contest model be adapted to solve timely task completion component in near real-time crowdsourcing tasks?
2. How does a change in reward spread and task threshold affect the worker's effort (number of annotations) and worker's output quality (quality score)?
3. How does a change in reward spread and task threshold affect the exit behaviour of workers (that is, how quickly do they exit)?
4. Can we predict when a worker would leave a crowdsourcing contest given their current performance?

7.5.2 Research Data

Our experiment dataset consisted of $7,600$ aggregated from four existing corpora from the literature. These datasets were from different time frames and had published gold standards, which we could use to perform quality checks and compute contest scores.

- ***The Ritter corpus*** by Ritter et al. (2011) which consists of $2,400$ tweets. The tweets were randomly sampled, however the sampling method and original dataset size are unknown. It is estimated that the tweets were harvested around September 2010.
- ***The Finin corpus*** by Finin et al. (2010) consists of 441 tweets which was the gold standard for a crowdsourcing annotation exercise. It is not stated how the corpus was created, however our investigation puts the corpus between August to September 2008.
- ***The MSM 2013 corpus***, the Making Sense of Microposts 2013 Concept Extraction Challenge dataset by Basave et al. (2013), which includes training, test, and gold data; for our experiments we used the gold subset comprising 1450 tweets.

- ***The Wordsmith corpus***, reported in one of our previously published works (Feyisetan et al., 2015a). From the corpus of six billion tweets, we sampled out $3,309$ English ones using *reservoir sampling* – a family of randomized algorithms for sampling k items from a list S of n items.

7.5.3 Worker Recruitment

We recruited our contestants from CrowdFlower, a marketplace for paid microtasks in which requesters posts tasks and crowd workers select tasks to work on. In order to achieve timely worker recruitment, we adopted a combination of strategies:

- We posted our tasks repeatedly in order to maintain visibility within the recent tasks view of workers;
- We created multiple tasks that pointed to our Wordsmith platform, but ensured that workers could connect only once by keeping track of the connection IP address;
- We attempted to recruit, on the average, 10 times the number of workers than we required ($\approx 1,000$ workers);
- We posted the tasks in bits as a work around for the scheduling mechanism which CrowdFlower uses in displaying unfinished tasks to new workers; and finally,
- We used an audio alert to notify workers once the requisite number of contestants had connected to the system.

Workers could see in real-time how many more contestants were required to connect before the task started. We also ensured that impatient workers were reconnected whenever they refreshed their screens, however, once the required number of workers were connected, no further connection was allowed.

7.5.4 Reward Spread

The reward spread represented the number of workers who were going to be paid for a given contest. A reward spread of 1 stands for a winner-takes-all condition, in which only the top worker gets paid. Our model rewarded each winner with the same payment, as opposed to other variants which take into account the ranking of the participants or implement some other form of reward sharing. As a result, in a reward spread of 10, the top-10 workers are paid *$x* each for the work they have carried out, while in a reward spread of 5, only the top-5 would have this benefit. We experimented with three different reward spreads:

- **Top 1**: 1 worker gets paid
- **Top 5**: 5 workers get paid
- **Top 10**: 10 workers get paid

7.5.5 Task Threshold

The task threshold represents the minimum number of tweets which contestants were required to annotate. We experimented with two task thresholds:

- **Low threshold** condition, contestants were required to exert the minimal amount of effort (equivalent to annotating 1 tweet) to qualify for the available reward(s)
- **high threshold** condition, contestants were asked to put in more effort, in our case 10 tweets i.e., 9 more tweets than in the low threshold condition.

No matter what the task threshold, if the worker was among the top contributors according to the reward spread (the first, in the top 5, or in the top 10, respectively), they would get paid the amount that was agreed as payment for the specific task (annotate one tweet; or ten).

7.5.6 Experimental Conditions

We carried out a within-subjects study, in which a number of workers were recruited from a large pool to participate in 6 different contests, taken into account the two parameters (reward spread and task threshold) discussed earlier. These were:

C1: pay top worker, at least one tweet per task;
C2: pay top worker, at least 10 tweet per task;
C3: pay top 5 workers, at least one tweet per task;
C4: pay top 5 workers, at least 10 tweet per task;
C5: pay top 10 workers, at least one tweet per task;
C6: pay top 10 workers, at least 10 tweet per task.

Within each experiment sub-condition, we recruited 100 contestants. Each contest had a payoff of \$0.10 for each prize payment, replicating Feyisetan et al. (2015a) and the experiments from Chapter 6.

7.6 Results

We recruited 100 microtask workers from CrowdFlower for each of the six experimental conditions. Table 7.1 lists how many workers carried out annotations on at least one tweet; as workers were allowed to skip tweets or exit the contest all together, in each of the six cases fewer than the 100 recruited workers actually ended up delivering on their tasks. In order to ascertain the consistency of the results, we repeated the experiments twice.

The findings confirmed the viability of the model for completing microtasks in near real-time and faster and cheaper than other approaches. In the following sections we will outline the effects of the two experimental parameters, reward spread and task threshold on delivery time, output accuracy, output volume, and contest exit behavior.

Experiment	Pay top 1		Pay top 5		Pay top 10	
	LT	HT	LT	HT	LT	HT
Participants	80	73	70	77	86	67

Table 7.1: Number of workers that carried out at least 1 annotation

7.6.1 Delivery Time

In Figure 7.3 we compare the time spent annotating an entity in the different contests vs. a baseline entity annotation task (on the same datasets) (Feyisetan et al., 2015a). In the baseline experiments, workers were not placed under any specific time constraint, however, inherent timely completion was required to receive their task compensation - therefore, they still had an incentive to complete the annotations without any delay. In experiments $C1$ to $C6$, workers needed on average 4.70 seconds to recognize and type an entity, vs. 9.70 seconds in the baseline experiment. This is equivalent to an annotation speed factor of $2x$ across all the contests. Workers spent an average of 5.6 seconds annotating one tweet (at 1.2 entities per tweet). The top delivery time was at 2.69 seconds per entity, equivalent to a speed factor of $3.6x$. Workers in the low threshold condition with '*pay top 5*' (condition $C3$) achieved the highest annotation speed at 4.33 seconds per entity.

Varying how many workers potentially received a reward did not yield a significant trend in how fast or slow workers carried out the annotation task. As noted earlier, the '*pay top 5*' experiments (conditions $C3$ and $C4$) resulted in the quickest annotations, followed by the '*pay top 1*' ($C1$ and $C2$), then the '*pay top 10*' workers ($C5$ and $C6$).

However, higher task thresholds did yield a trend in annotation delivery. When workers were required to hit a higher threshold before they could potentially get paid, they were slightly faster - this was consistent across all the experiment conditions.

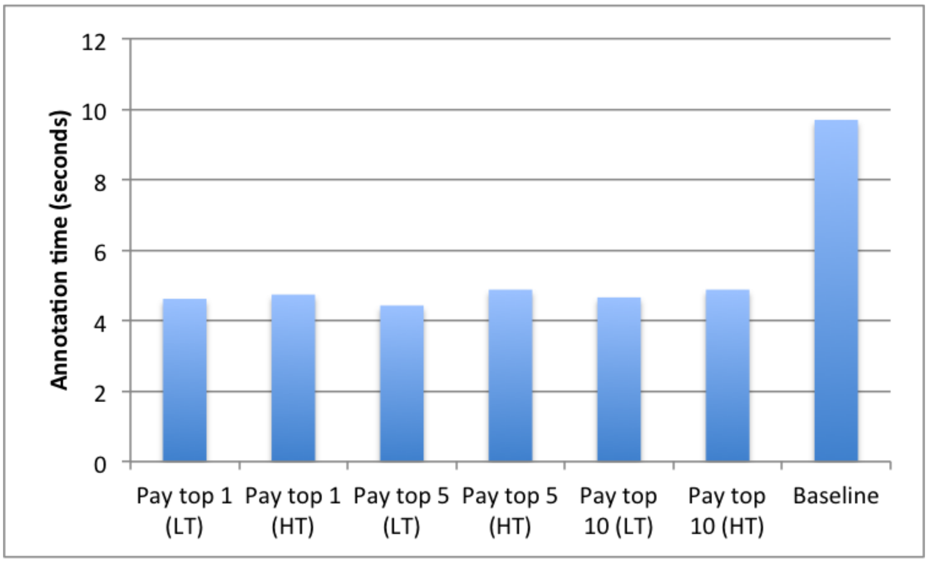


Figure 7.3: Average annotation time per entity

7.6.2 Annotation Quality

Faster results did not come without quality compromises. Figures 7.4 and 7.5 present a fine grained summary, evaluating the annotation quality with respect to the associated dataset gold standards. The experiment quality results are illustrated side-by-side with the results from the baseline experiments. We look at precision, recall, and F1 scores for the baseline, and each experiment condition $C1$ to $C6$ for each type of entity.

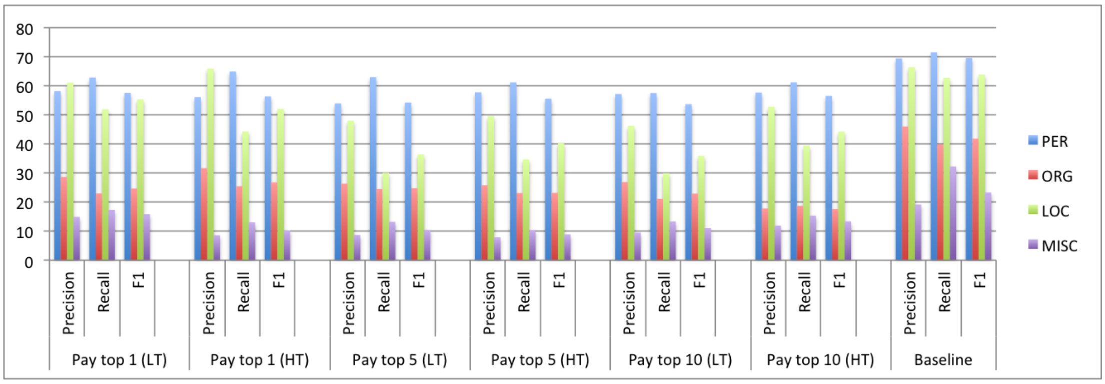


Figure 7.4: Annotation accuracy compared with baseline highlighting precision, recall and F1 score breakdown

Figure 7.4 reveals that the experiments which rewarded only the top worker (in the low threshold condition $C1$) produced the F1 score with the highest aggregated value. As the reward spread increased to conditions where 5 workers, and 10 workers were eligible to be paid, the quality dropped and then flat-lined across the other four conditions ($C4$ to $C6$).

Figure 7.4 further suggests that varying the task threshold only effects a significant change in the condition where in a winner-takes-it-all condition. The aggregated F1

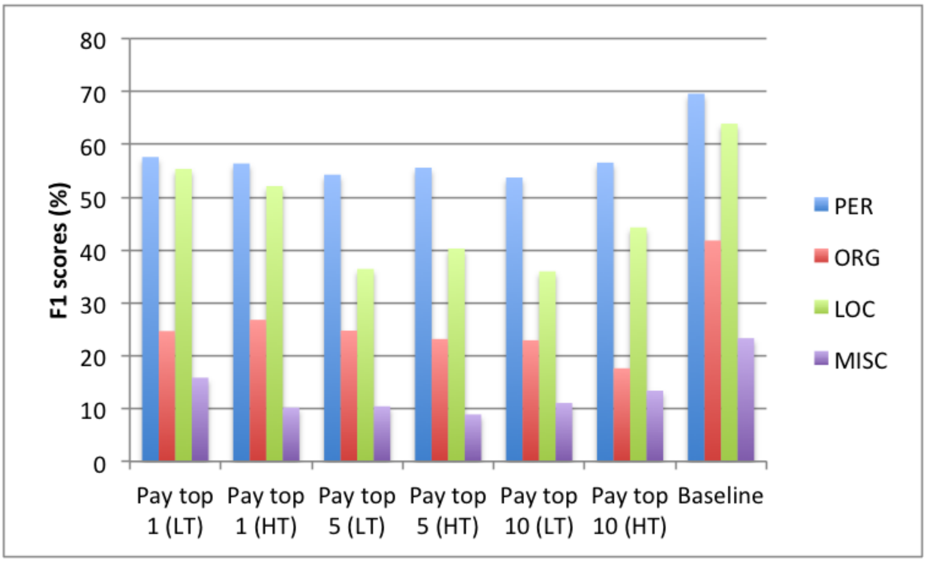
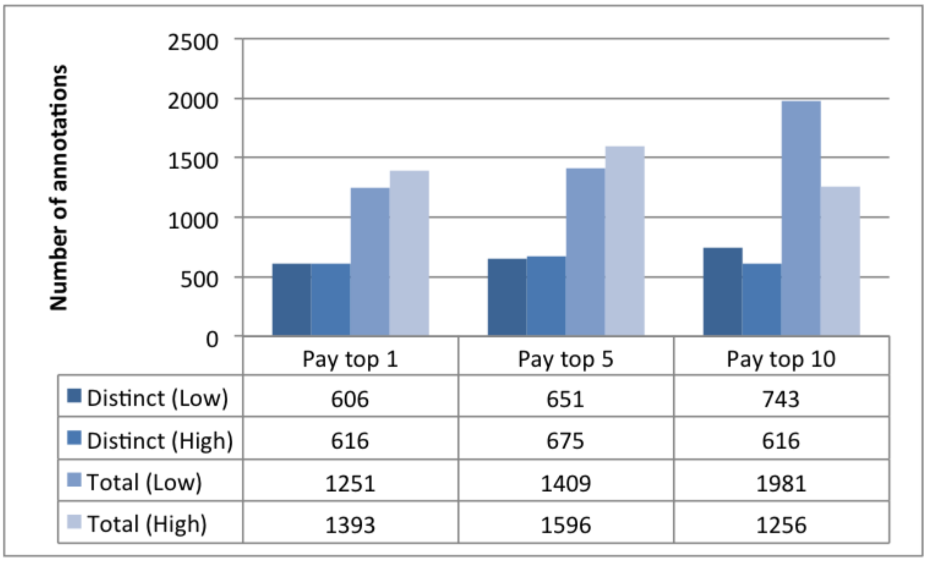


FIGURE 7.5: Annotation accuracy compared with baseline highlighting only F1 scores

score was higher for low threshold tasks. In the subsequent experiments, varying the task threshold did not result in an improvement or reduction in the average F1 score.

7.6.3 Task Output

Figures 7.6 to 7.11 illustrates how much effort was exerted by the workers across experimental settings. Figure 7.6 shows the total output numbers, i.e., the distinct and total number of annotations carried out for each experiment, while Figure 7.7 presents the average number as a function of the number of workers who carried out actual annotations. Figure 7.8 then displays the output by the top annotator in each of the experiments, while Figure 7.9 is about the total annotations by the top 10 workers in all experiments.



	Pay top 1	Pay top 5	Pay top 10
■ Distinct (Low)	606	651	743
■ Distinct (High)	616	675	616
■ Total (Low)	1251	1409	1981
■ Total (High)	1393	1596	1256

FIGURE 7.6: Total distinct annotations by workers

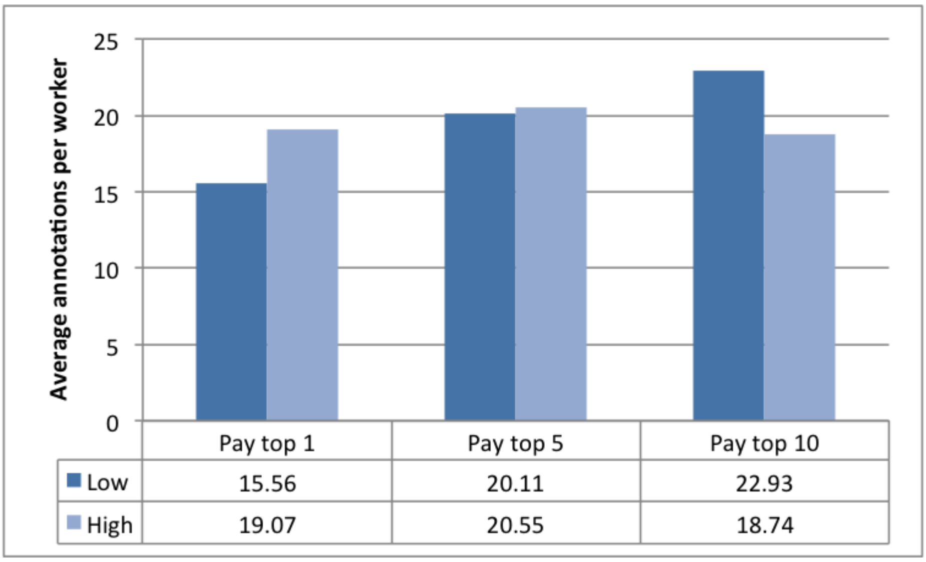


Figure 7.7: Average annotations per worker

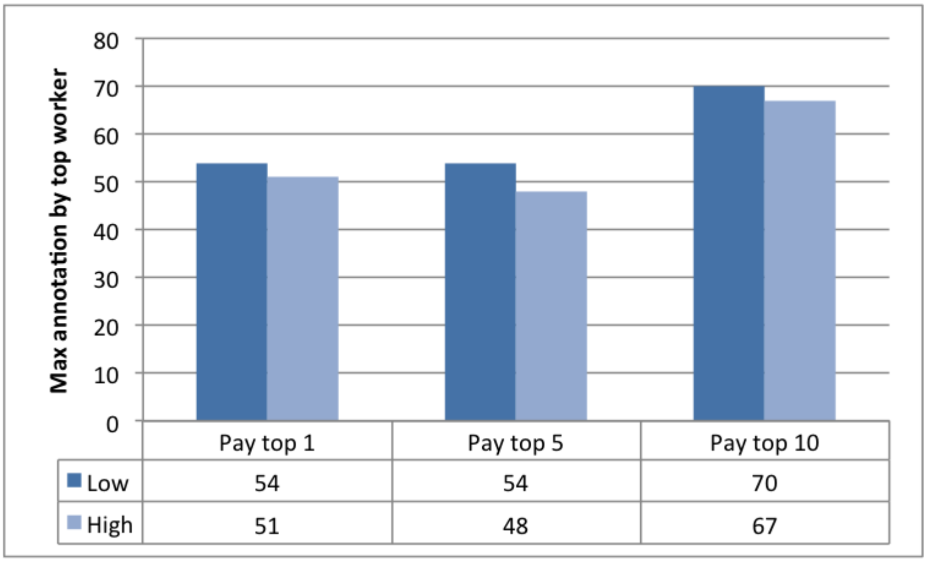


Figure 7.8: Number of annotation by top worker

The results demonstrate that being willing to reward more workers increases the overall efforts exerted by the crowd. In Figure 7.6, changing the reward spread from 1 to 5 to 10 led to more annotations overall. Figure 7.7 presents a similar trend, with the average number of annotations increasing with the reward spread. We also note an increase in the number of annotations by the top workers from the first two sets of experiments, where the output was fairly constant, to the third condition (with 10 workers were eligible for payment), where there was a significant increase in the number of tasks performed (Figure 7.8). Figure 7.9 exhibits a similar pattern with more output of the top performers across the experiment conditions.

Varying the task thresholds within each reward spread category presented, however, a slightly different set of results. From Figure 7.6 and Figure 7.7, we see that the total

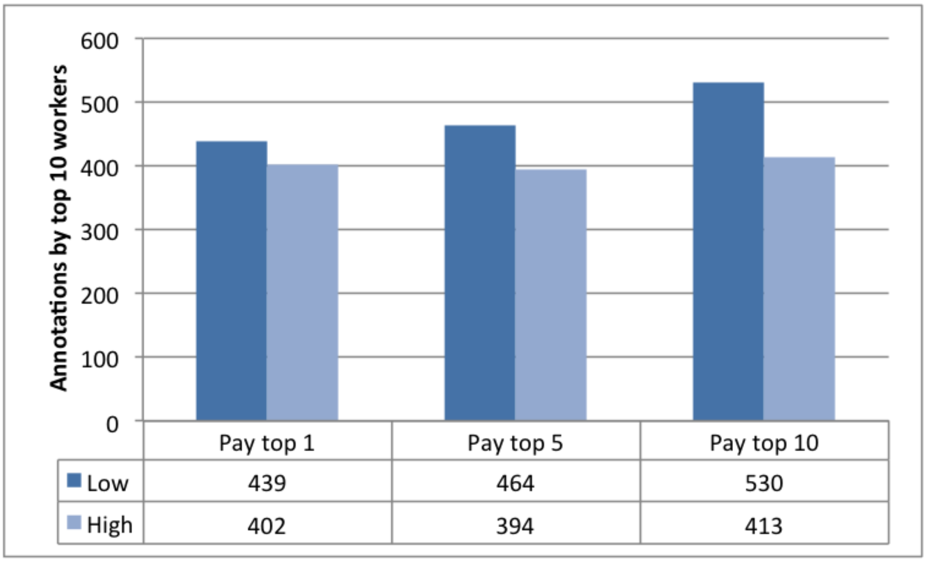


Figure 7.9: Annotations by top 10 workers

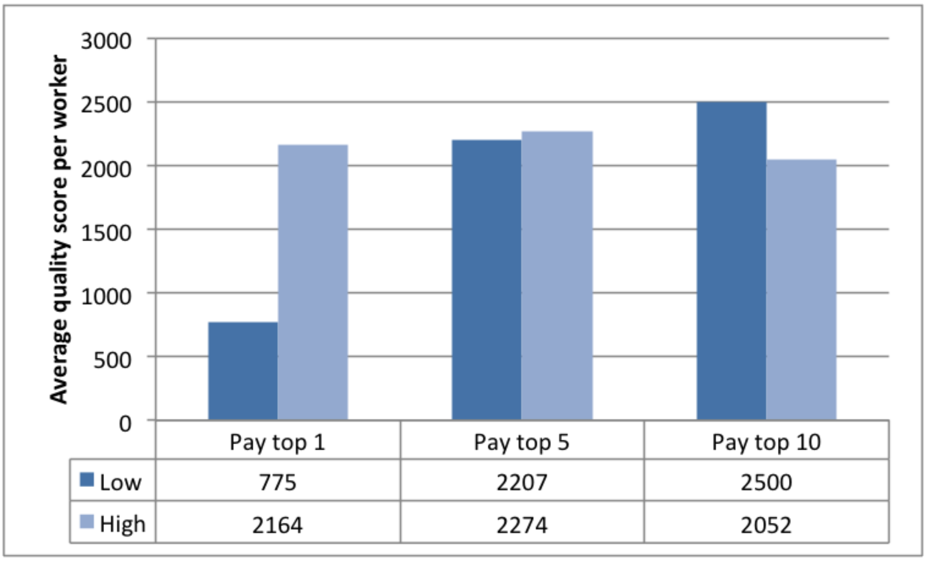
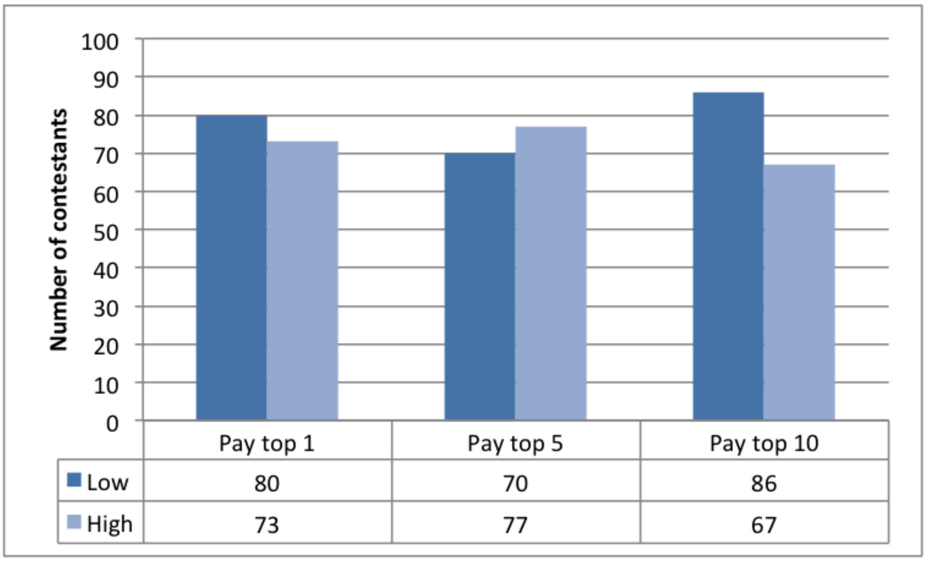


Figure 7.10: Average quality score per worker

and average number of annotations go up from the low threshold, where workers were required to mark entities in just one tweet, to the high threshold, where they had to complete 10 tweets. This is unsurprising as the high threshold conditions required more effort to receive a potential payoff. In $C5$ and $C6$, however, there was a significant drop in the values of these metrics (despite these experiments leading to a high number of tasks completed by the top performers). We re-ran the experiments to check for any external factors that might have caused this effect and the same results were observed. From Figure 7.8 and Figure 7.9, we observe that increasing the task threshold (within the same reward spread condition) consistently leads to a reduction in the output of the top performers i.e., even though the overall effort exerted by all workers in higher, top contributors engage less.



	Pay top 1	Pay top 5	Pay top 10
Low	80	70	86
High	73	77	67

FIGURE 7.11: Number of contestants

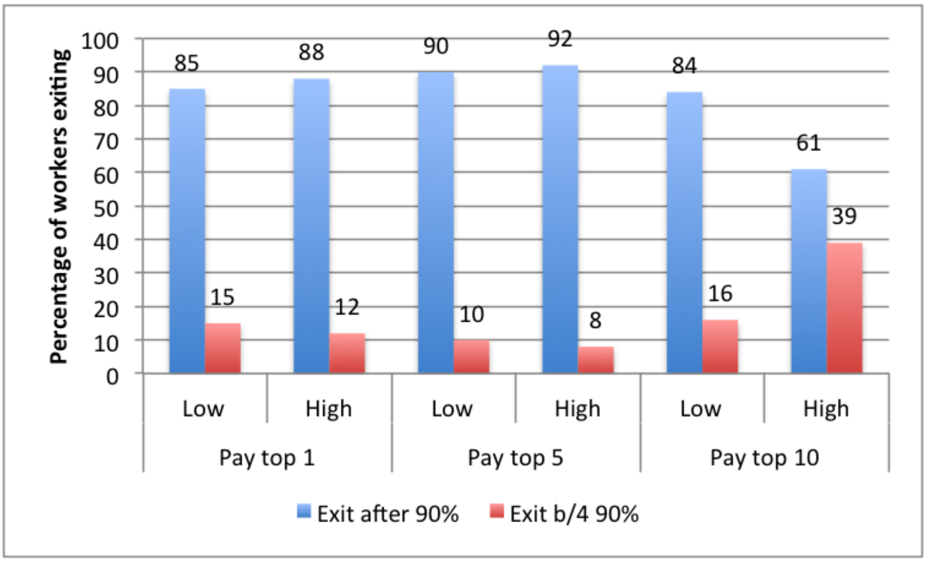


FIGURE 7.12: Exit distribution

7.6.4 Contest Exit Behaviour

Only a subset of contestants received a monetary payoff. The longer a worker engaged with a task, the more utility they potentially lost given their ranking relative to the reward spread. Some workers therefore opted to leave the contest. Figure 7.12 presents the exit behaviour of workers in all experiments. The results imply that the beginning of the contests sees fewer exits with most of the contestants choosing to continue for up to 90% of the total time period. The exit rates increase towards the end. This behaviour would be peculiar to microtask contests unlike much longitudinal contests such as presidential elections (studied by Norrander (2006)), in which most contenders exit at the beginning of the race leading up to much fewer participants at the final elections. Figure 7.12 illustrates this phenomenon by presenting the percentage of workers that

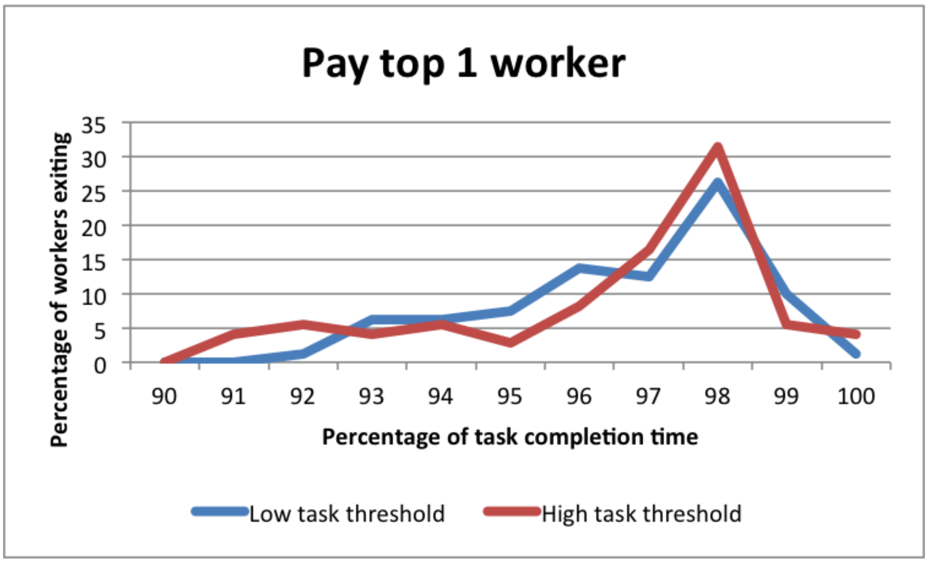


FIGURE 7.13: Pay top 1 worker

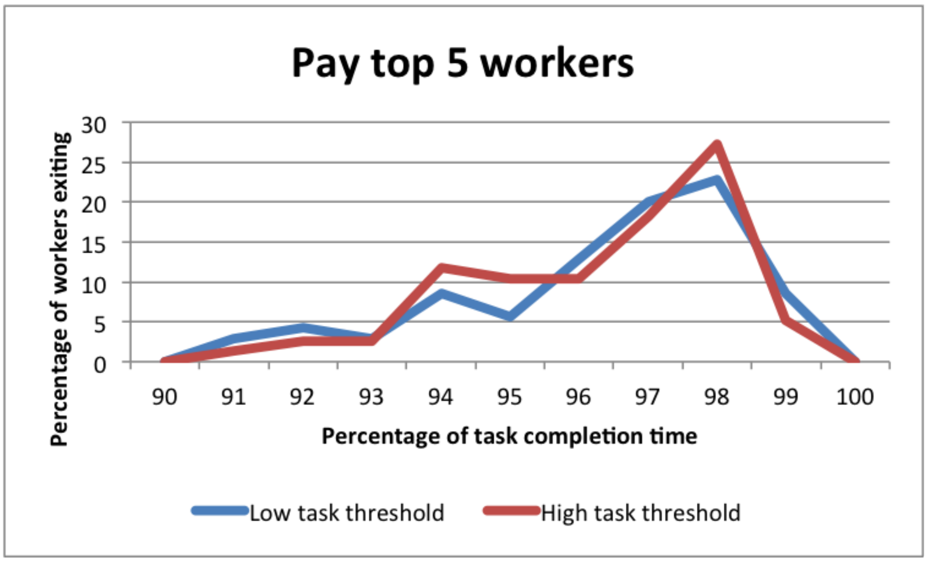


FIGURE 7.14: Pay top 5 worker

exit the contests before and after the 90% time cutoff. This aspect is presented in Figure 7.12 for the six experiments. Figure 7.13, 7.14, and 7.15 on the other hand focus on the last moments of the contest - the final 10% time stretch. The figures present the exit distribution, i.e., the number of workers quitting the contest at various discrete time spots leading to the end of the task.

Figure 7.12 reveals a slight continuous increase in the number of workers staying past the the 90% threshold from conditions $C1$ and $C2$, to $C3$ and $C4$. Consequently, in these four conditions, the percentage of workers exiting the contest early reduced gradually. The final two experiment conditions $C5$ and $C6$ present an opposing result, with a decline in the total number of workers making it to the 90% threshold. As has been seen from the results discussed so far, the top performers in these conditions (who stay up until the

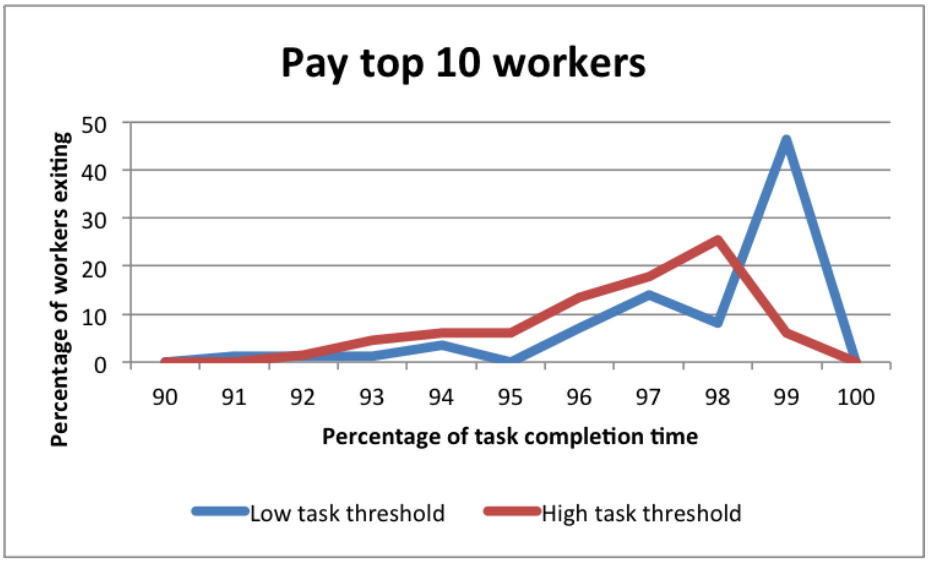


FIGURE 7.15: Pay top 10 worker

end), go on to annotate more tweets than the preceding scenarios. Similarly, looking at the details of exit behaviour in conditions $C1$ to $C4$ in Figure 7.13 and 7.14, we observe similar exit patterns leading up to a sharp mass exodus at about the same time. The area under the curves in the two figures is proportional to the numbers illustrated in the 'exit after 90%' bars in Figure 7.12. Contrasting the two result sets, we note that in Figure 7.14 (conditions $C3$ and $C4$), there is a more gradual buildup of workers exiting, up to the peak (98% contest time), when over 25% of workers pulled out. In Figure 7.13 (conditions $C1$ and $C2$), the buildup occurs slightly later leading to a larger exit of over 30% of workers at the 98% peak. From Figure 7.15 (conditions $C5$ and $C6$), we see more workers staying longer all the way almost to the end i.e., despite having a higher initial attrition rate before the 90% threshold, more workers stayed on in attempts to qualify for the top 10 spots.

The results of $C1$ through to $C4$ displayed in Figures 7.13 and 7.14 demonstrates that increasing the task threshold results in workers remaining longer in the contest, which is also due to an increased task baseline. This is seen from the area under the curves and the corresponding result in Figure 7.12. As with our previous findings, the inverse was the case in the final two experiment conditions ($C5$ and $C6$). Increasing the task threshold and potentially paying more people not only led to fewer participation from the outset (67 workers in $C6$), but it also led to more workers exiting the contest sooner.

7.6.5 Predicting Contest Exit

We evaluated our predictive model by carrying out a 6-fold cross-validation on our 6 experiment result sets. The training data consisted of the state of each worker at every time point (such as number of tweets annotated so far, etc., see section 7.4). All the result sets were randomly split into 6 parts of equal sizes; 5 of the sample parts were then used to train the model and the last to evaluate. This process was then repeated 6 times. This gave scores of 71.95, 71.96, 71.96, 71.96, 72.22, 72.22 for an average of 72.04%.

Correctly predicting 7 out of 10 exits, especially at the early stage can potentially lead to significant increase in task output by applying appropriate furtherance incentives (introduced previously in Chapter 6 and expanded later on in Chapter 8). We can even increase the accuracy of the prediction by creating a finer grained model around the worker e.g., taking each time state and examining the worker's annotation rate up to that point.

7.7 Discussion

In this section, we revisit the main findings in the context of the research hypotheses introduced earlier, and discuss their implications for this line of work and for microtask crowdsourcing in general.

7.7.1 Towards Real-Time Annotations

The results presented lend support to our hypothesis that the microtask contest model leads to timely task completion without the associated overhead costs. Our results indicate an increase in speed by an average factor of $2x$, and up to $3x$ more among the top performers. These metrics could be used to inform decisions on the number of workers that would be needed to annotate the entire dataset of $7,600$ tweets in real-time - taking into consideration the total number of non-unique annotations e.g., in experiment $C5$, workers annotated 1981 tweets, equivalent to 384 workers required to make one pass at the entire stream.[2] Our experiments were carried out on a stream of 20 tweets per second; the live Twitter stream is currently estimated at $6,000$ tweets per second[3] or $2,400$ English tweets per second (40% of the full stream) (Feyisetan et al., 2014). Annotating 10% of the live stream could be potentially carried out by designing a contest for $\approx 4,600$ workers i.e.,

$$4,600\ workers \quad \leftarrow \quad 384\ workers\ * \ \frac{10\%\ of\ \ 2,400\ tweets}{20\ tweets\ per\ second} \tag{7.7}$$

[2] This figure includes consideration of worker exit.

[3] http://www.internetlivestats.com/twitter-statistics/

7.7.2 Reward Spreads and Task Thresholds

We investigated the interplay between the reward spread, task threshold, and crowd behavior. Allowing more workers to be eligible for payment improved the overall performance. With more winning spots came an increase in the total and average task output by all participants. Having a higher reward spread also ensured that workers stayed in the context and did not drop out early, thus leading to more tasks completed. This further meant that more effort was required to achieve one of the top spots in the ranking, which would receive a payment.

Delivery times remained fairly stable across the experiments, hence an important factor required was reducing the worker attrition in the contest, which was achieved by increasing the reward spread. High threshold tasks improved the amount of work produced; even though the top performers weighed their contributions more carefully, the workers at the lower end of the leaderboards evened out this shortfall by doing more annotations.

However, indefinitely increasing the reward spread would defeat the purpose of adopting a contest model, converging towards a traditional microtask system. Our results suggest that the linear result growth begins to break down as expected at some point. It is important to note that the motivation of crowd workers covers a wide spectrum of intrinsic and extrinsic factors, hence, having a wider reward spread (and a higher task threshold as seen in $C6$) led to a plummet in task output (over repeated experiment runs). An investigation into the discussion forums indicated that this experiment was probably less challenging – as stated by one worker, '*... to get into the top 10 is not too difficult ...*' – and hence might not have been as attractive to top performers. Understanding where to draw the line would be the subject of further empirical studies paired up with theoretical analysis.

7.7.3 Payments and Ethical Considerations

Crowd worker motivation remains a constant research area in understanding why people partake in microtasks. This is essential in order to design systems which are fair and rewarding to the workers. A requester would always seek to minimise cost, however, a worker's complete range of motivations is not yet fully understood. Our investigations revealed that the task model was relatively well received: in one of the baseline experiments, 87 workers rated the payment of $0.10 as 4 out of 5, while in one of our contests, 49 workers rated it as 3.5 on the same scale despite only 9 of them receiving an actual payout. Furthermore, workers seemed to be eager to return – as stated in a crowd discussion forum, one of the participants posted '*Hit the top* 10 *today. I will hunt this problem again*', while another one felt let down when they couldn't be among the 100 contestants: '*Yesterday I came across this, but [they] recruited 100 people, [I] was not allowed to play*'.

On task payment, we favoured a higher than average payment of \$0.10 as against the annotation averages reported by Difallah et al. (2015), however, it would be interesting to see the effects of increased payments on the results. In a related set of experiments we noted that raising the reward to as much as \$0.25 rather created an anchoring effect than a trend in results. We would like to investigate this further and also analyse the effect of task payment ordering. A greater understanding of worker intrinsic motivations would help in the design of better payment schemes, wrapped around task models that workers find inherently engaging and rewarding; and lend insights into the ongoing debate on ethical and fair crowdsourcing (Irani and Silberman, 2013).

7.7.4 Limitations

In this chapter we did not focus on advanced task allocation mechanisms, however our results can be extrapolated to what the results might look like. The experimental findings presented the distinct and total number of annotations by workers, which both give an idea of the number of workers that would be required to attain complete task coverage, within the given time constraint using an optimal task allocation strategy. Furthermore, the experiments were run on a named entity recognition task. While we tried not to focus on task specificity and present a model that could be generalised to different task scenarios, it would be interesting to explore how this model would perform in other task settings.

7.8 More on Furtherance Incentives

In Chapter 6, we introduced the concept of furtherance incentives and how they could potentially be used to design better crowdsourcing workflows by engaging workers who are about to quit the task with sub-tasks that they are good at solving. Earlier on also in Section 7.6.5, we have similarly alluded to the potential of using furtherance incentives to improve real-time crowdsourcing contests by predicting workers who are about to quit and pre-emptively engaging them.

We therefore broadened our discourse on the potentials of furtherance incentives in this chapter. The keys to effectively deploying furtherance incentives come in two folds:

1. we should be able to detect when workers are about to quit a task. This has been identified as a possibility in this chapter by employing predictive analytic capabilities which observe worker contest and exit patterns to create a picture of when workers would likely quit the task; and

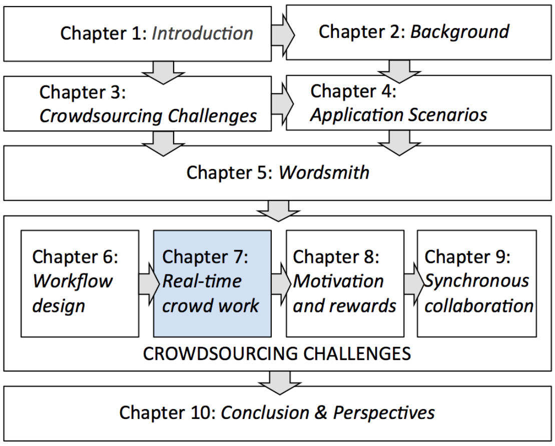
2. we should have an idea of ‘tipping point’ incentives which can effectively keep a worker engaged i.e. switch a disinterested worker to a sub-task (or set of sub-tasks) that the worker would be more likely to engage with.

In the previous chapter, we presented possible furtherance incentives in the way of tweets which had been demonstrated to be quickly and correctly annotated by the crowd i.e. fulfilling step (2) above; while in this chapter, we provided a way to detect or predict when a worker is about to quit a task, therefore addressing point (1) above. In the next chapter, i.e., Chapter 8, while discussing the challenge of motivations and rewards in crowdsourcing, we would bring these two components together thereby painting a more rounded picture of how to deploy furtherance incentives. We would discuss different incentive types in a gamified paid microtask setting and present analysis on the effectiveness of using the various incentives as furtherance incentives.

7.9 Conclusion

We demonstrated that designing tasks as contests can speed up the completion factor by an average of $2x$, making them suitable for real-time crowdsourcing. Furthermore, we demonstrated that increasing the reward spread and task threshold increases the overall task output up to a certain point after which, the result size begins to decline. Our results illustrate that increasing the reward spread prevents early exit of workers. We also reported a rather positive impression from task workers based on the satisfaction scores (even of unpaid workers) and forum posts which suggests a motivation factor beyond the baseline payment. These results could be used to inform better real-time crowdsourcing systems within budget constraints without sacrificing the intrinsic benefits workers might derive from the platforms.

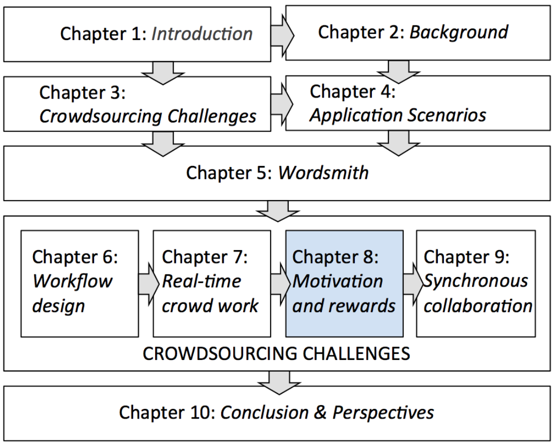
7.10 Summary


Chapter 1: *Introduction*
Chapter 2: *Background*
Chapter 3: *Crowdsourcing Challenges*
Chapter 4: *Application Scenarios*
Chapter 5: *Wordsmith*
Chapter 6: *Workflow design*
Chapter 7: *Real-time crowd work*
Chapter 8: *Motivation and rewards*
Chapter 9: *Synchronous collaboration*
CROWDSOURCING CHALLENGES
Chapter 10: *Conclusion & Perspectives*

The results in this chapter illustrated the viability of applying a contest model to carrying out crowdsourcing tasks with time constraints. We presented a predictive model that was used to suggest when a worker wanted to quit the task. Afterwards, we gave empirical results to show crowdsourcing settings that yield useful results under time constraints using the contest model. Following from the previous chapter, we expanded our understanding of furtherance which leads squarely to the central theme of the next chapter.

Chapter 8

Motivation and Rewards


Chapter 1: Introduction
Chapter 2: Background
Chapter 3: Crowdsourcing Challenges
Chapter 4: Application Scenarios
Chapter 5: Wordsmith
Chapter 6: Workflow design
Chapter 7: Real-time crowd work
Chapter 8: Motivation and rewards
Chapter 9: Synchronous collaboration
CROWDSOURCING CHALLENGES
Chapter 10: Conclusion & Perspectives

In this chapter, we address the challenge of motivation and rewards in paid microtask crowdsourcing; building upon, and coming full circle on the concept of furtherance incentives. This chapter examines the potential of adding gamification to microtask interfaces as a means of improving both worker engagement and effectiveness. It also defines a predictive model for estimating the most appropriate furtherance incentive for individual workers, based on their previous contributions. This allows us to build a personalised game experience, with gains seen on the volume and quality of work completed.

This chapter is adapted from earlier published work [1] titled 'Improving Paid Microtasks through Gamification and Adaptive Furtherance Incentives'.

8.1 Overview

We run a series of experiments in image labelling, and analyse worker behaviour in terms of number of images completed, quality of annotations compared against a golden standard, as well as monetary and game-specific rewards. Each experiment studies these parameters in two settings: one based on a state-of-the-art, non-gamified 'job' on CrowdFlower (i.e., the unit of work on this platform); and another one using an alternative

[1] This chapter is adapted from work that appeared at WWW 2015 Feyisetan et al. (2015b)

interface incorporating several game elements. The second setting uses CrowdFlower as well, but only to seek contributors; it offers the same reward for the same amount of work as the baseline task, but points to an external page where the gamified version of the task is deployed. More specifically, in the second condition, CrowdFlower workers are asked to engage with Wordsmith.

Our basic hypothesis is that by designing a playful interface for the image labelling task - as opposed to the functional style common to most microtask platforms - we will encourage workers to engage with the task more, independently of the actual monetary reward. This hypothesis was confirmed by our findings, which revealed better accuracy (an improvement of almost 10% compared to the baseline condition) and significantly lower costs per annotated image (5,708 unique labels collected via the game vs. 111 unique labels contributed through equivalent, non-gamified microtasks, see Experiment 1 in Section 8.4). We tested this hypothesis on different variations of image labelling tasks, in which we increased the complexity of the task and adjusted the prices accordingly, observing a similar trend.

Then we looked into the impact of different game elements and related incentives on the behaviour of the workers, following the *SAPS framework* (Status, Access, Power, Stuff) presented by Zichermann and Cunningham (2011). Besides studying how people responded to the primary gamification components (leaderboard, levels, points, and badges), we also introduced a sociality aspect, which was originally missing from the game and allowed contributors to view each other's achievements. Our final incentive dimension was additional cash payment for more work, once the goals of the initial task had been achieved.

These new experiments produced clear evidence of the positive effects of game mechanics on both task performance and crowd engagement; up to five times more unique labels were generated while preserving a comparable level of accuracy (see Experiment 2 in Section 8.4). This is true particularly for sociality features, which are largely absent on microtask platforms.

Following these insights, we went on to create a predictive model that estimates the most suitable set of incentives for individual workers, based on their previous contributions. This allowed us to build a personalized game experience, with positive results on the volume and quality of work completed. With this model, we were able to obtain 19% more concordant image keywords (4849 vs. 4091) while maintaining the same average pair-wise agreement score. We also recorded a significant uptake in image tagging (with 77% of players tagging an extra image when confronted with targeted incentives against 27% in the experiments using a randomly selected incentive).

Overall, the results of these experiments shed light on possible design improvements of paid microtasks environments in order to achieve better task performance and make the overall experience more fair and rewarding for the workers. While we are not necessarily arguing for a fully-fledged gamification of microtask platforms, considering specific game mechanics (Zichermann and Cunningham, 2011), or in fact, any social design features that are widely discussed on online communities literature (Kraut et al., 2012) is worth further investigations. This is important not just for purely utilitarian motives on the side of the task requesters, but also in the context of the ongoing debate on ethical and fair crowdsourcing (Irani and Silberman, 2013).

Previous work has approached such aspects through studies of crowd motivations (Kaufmann et al., 2011), discussing the rich repertoire of extrinsic and intrinsic reasons that drive people to contribute to microtask projects. Our experiments quantify some of these insights. We deliberately chose a task that is well-known in the crowdsourcing literature, as we were aiming for task-independent findings, which were only minimally influenced by interface or quality control aspects. For the same reasons, we opted for average market prices to reward participation; lower pays would have been less attractive (and unfair) for workers, higher ones might have appealed to people who were primarily financially incentivised. We believe more research needs to be done to build microtask platforms that reflect and support the values and motivations of the crowd as an integral part of their functionality. Our experiments give evidence that such efforts could be beneficial both workers, and for requesters.

8.2 Model

In this section, we introduce a high level overview of our approach to crowdsource image labelling. We present our microtask design model and strategies for undertaking crowd work. This involves the use of an external platform, CrowdFlower, and our bespoke game *Wordsmith*.

8.2.1 Task

We now describe our model for maximising the output from crowd assigned tasks while maintaining quality. Each HIT (Human Intelligence Task) consists of n images, $x = \{x_1, ..., x_n\}$, which can each be described by a set of m keywords $k = \{k_1, ..., k_m\}$, where m is a large unknown number. Each task seeks to capture new keywords that correctly describe each image.

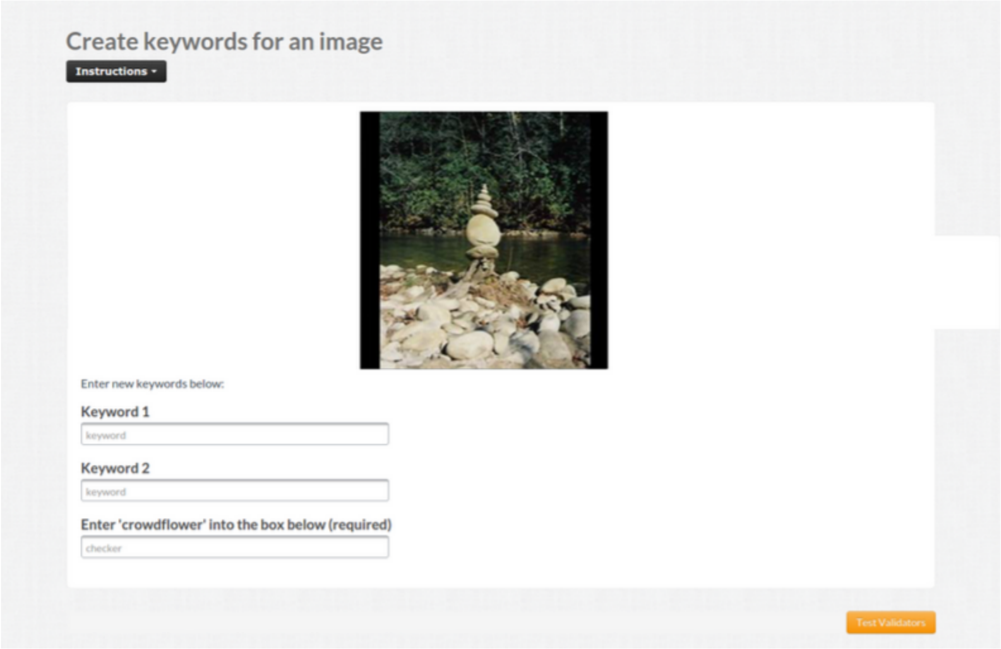
Requester. The requester desires to have as many image annotations as possible, without compromising on the quality of the describing keywords. The requester requires the help of human agents to carry out the tasks.

Strategy. We define two requester strategies for presenting tasks. The *crowd* strategy relies on traditional crowdsourcing techniques in a standard 'image field - text fields' layout. The *game* strategy employs game mechanisms, and a game based interface to capture keywords. Our nomenclature defines human agents in the crowd strategy as *workers*, and those in the game as *players*.

Crowd → Worker. Each worker provides judgement on a task by assigning m keywords $\{k_1, ..., k_m\}$ to n images $\{x_1, ..., x_n\}$ in a traditional crowdsourcing system. We used CrowdFlower as our crowdsourcing platform, presenting each task using the standard image annotation template provided. In this strategy, n and m are defined and fixed by the requester.

Game → Player. Each player provides judgements on a task in a fashion similar to workers in the *crowd* strategy. However, in the *game* strategy, players can tag a variable number of images as they progress through more levels.

Quality. Is defined by consensus. The number of keywords matching a quasi-gold standard bank, gives an overview of the quality of annotations. This was extended to also cover consensual annotations within workers and players - as this suggests probable new keywords for the image.


Create keywords for an image

Enter new keywords below:

Keyword 1

Keyword 2

Enter 'crowdflower' into the box below (required)

Figure 8.1: CrowdFlower Task Interface

8.2.2 Worker Recruitment

We sourced all our human agents from CrowdFlower. For each experiment, we created 2 jobs which channelled task resources to the crowd strategy and game strategy. We used identical settings for each experiment set, consisting of the following parameters:

1. Geography - limited to the top 15 English speaking countries, and the top CrowdFlower contributor countries.
2. Skills - we chose *Level 2 Contributors*, which account for 36% of monthly judgements.
3. Judgements - 3 per unit, which meant each image would be annotated by at least 3 human agents.
4. Behaviour - each human agent was paid for 1 task, i.e., paid to tag m images, with n keywords, each as determined by the requester. For this, CrowdFlower tracks the IPs and aliases created by the agents.
5. Reward/Time Limits - reward payment and completion time limits were experimentally set as described later.

8.2.3 Game Design

Apart from the baseline gamification elements, additional feedback mechanisms consisting of information were provided to players on their progress and current standing in the game. Both social and non-social feedback elements were added to Wordsmith as follows:

1. Leaderboard - showed the hourly scores and level of the top 5 players as opposed to an all time leaderboard which might discourage newer joiners.
2. Badges - were awarded based on the number of images tagged. The first set of badges are listed below:
   - Take Off Badge - tag 1 image
   - Second Shot Badge - tag 2 images
   - Steady Progress Badge - tag 5 images
   - Premier Cup Badge - tag 8 images
   - Premier Crown Badge - tag 13 images
3. Feedback Alerts - informed a player how a bonus point or badge was attained and how it can be re-attained. The alert messages displayed were as follows:
   - Badges - You are close to a badge ... Just a few more images and you will unlock an exciting new badge - you're almost there.
   - Bonus points - You earned bonus points! Typing in the right words earn you either single, double and triple bonus points
   - Treasure points - You earned treasure points! Earning multiple bonus points (10, 20, 30) stack up to earn you extra treasure points

4. Bonus Points - were awarded when players submitted keywords that matched 1, 2, or 3 known images tags.
   - single point - match 1 gold label
   - double points - match 2 gold labels
   - triple points - match 3 gold labels
5. Treasure Points - were awarded when players got multiples of 10 bonus points. There were 9 levels of treasure points; the first 3 are shown below:
   - bronze points - match 10 gold labels
   - silver points - match 20 gold labels
   - gold points - match 30 gold labels
6. Activities Widget - displayed in real-time, what other players were doing in the game. The categories of messages displayed were as follows:
   - Has unlocked - a badge, an avatar
   - Has earned - bonus points or treasure points
   - Climbed up - the leaderboard
   - Is now on - the global leaderboard
7. Levels - the game consisted of a total of 9 levels from Newbie to Wordsmith. A player's level advancement was a function of how many images were tagged.

Level name	Time allowed	Unit points	Min points	Max points	Reserved words	Required labels
Newbie	120	250	0	0	0	2
Novice	100	260	1	10	2	2
Competent	80	275	11	99	2	3
Master	70	295	100	199	3	3
Champion	65	320	200	299	3	3
Maestro	60	350	300	499	4	3
Commander	55	385	500	999	4	4
Grand Duke	50	425	1000	1499	5	5
Wordsmith	45	470	2000	∞	5	6

Table 8.1: Level Design

8.2.4 Furtherance incentives

In Experiment 2, we introduce '*furtherance incentives*' to Conditions 3 and 4, which is a reward or concession presented to a player when they attempt to exit the game to induce them to stay and play more levels. We selected our incentives based on the SAPS framework presented by Zichermann and Cunningham (2011).

In our experiment, we expanded *Status* from SAPS to encompass the 3 game status elements mentioned by Zichermann and Cunningham (2011), i.e., badges, leaderboard and levels. We interpreted the SAPS incentives as a popup messages presented to the player at the point of attempted exit. Each incentive began with the message: Would you like to tag the next '*target number*' images? If the player had tagged less than 21 images, the '*target number*' of additional images was 5, otherwise it was 11.

The specific messages appended to each incentive is as follows:

- **Badges** - You would automatically be rewarded with The 'Ultimate' Badge. Get upgraded to a shiny new avatar
- **Leaderboard** - You would automatically be advanced on The Global Leaderboard. Get seen globally on the leaderboard
- **Levels** - You would automatically be advanced to The Next Level. Advance to the next level.
- **Access** - You would be given quicker access to Treasure Points. Get more treasure in half the time.
- **Power** - You would be rewarded with the power to View Other Players Tags. Power to see other players image tags.
- **Stuff** - You would be rewarded with a bonus of 5 cents extra. More cash for your effort.

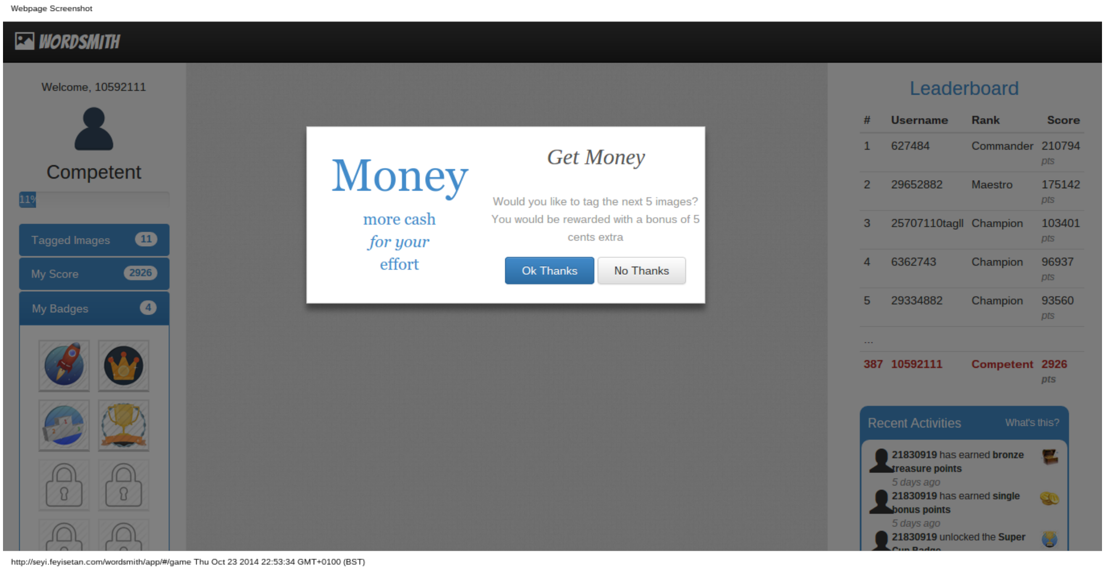
At each point of attempted exit, a player was shown one of the 6 furtherance incentives

$$V = \{badges, leaderboard, levels, access, power, money\}$$

The choice of the incentive to be shown was decided by drawing a random variable $V \backsim U([0,1])$

At the moment of incentive offer, we record the incentive offered, the requested target number, the number of images tagged so far (as start_tags and end_tags) and the current timestamp. We then recorded the player's game state after the incentive was presented i.e., the player could ignore the incentive and exit the game ($state{=}out$), or the player could go on playing the game ($state{=}in$).

If the player remains in $state = in$, we keep track of game play (updating the number of images tagged as end_tags) until the player has tagged an additional 'target images'. At this point, the offered incentive is activated. Players can then transition into state in or



Webpage Screenshot

Get Money

Would you like to tag the next 5 images? You would be rewarded with a bonus of 5 cents extra

#	Username	Rank	Score
1	627484	Commander	210794 pts
2	29652882	Maestro	175142 pts
3	25707110tagll	Champion	103401 pts
4	6362743	Champion	96937 pts
5	29334882	Champion	93560 pts
...			
387	10592111	Competent	2926 pts

21830919 has earned **bronze treasure points**
5 days ago
21830919 has earned **single bonus points**
5 days ago
21830919 unlocked the **Super Cup Badge**

http://seyi.feyisetan.com/wordsmith/app/#/game Thu Oct 23 2014 22:53:34 GMT+0100 (BST)

Figure 8.2: Furtherance incentive presented when a player attempts to exit Wordsmith

out. If a player attempts to exit the game at this point, we do not show any furtherance incentive. However, if the player remains in the game, we continue to keep track of the number of images tagged, and therefore update the value for end_tags.

8.2.5 Adjusting Incentives Probabilistically

In the final condition (Experiment 2 C4), we posit that certain furtherance incentives are more effective at different stages of gameplay. To test this hypothesis, we computed a probabilistic model that estimates, at every potential game exit point, the incentive that would maximize the probability of the player remaining in the game. To do this, our probabilsitic model computes *a priori* state transitions at previous attempted exit points, to predict what incentive a player would accept given the number of images they have tagged. This model is computed from:

1. ***the prior probability*** of the incentive given the incentive distribution obtained from the results of the random incentives condition (Experiment 4 Condition 3)

2. ***the likelihood probability*** of the player remaining in the game after tagging the current number of images given a certain offered incentive and

3. ***the likelihood probability*** of the player remaining in the game after tagging a set of images (defined over a numeric range), given a certain offered incentive.

The details of the reasoning approach is detailed in the next few sections.

Our probabilistic reasoning approach to computing the maximum a posteriori incentive given our selected feature (the number of tagged images) is similar to the method of determining the correctness of worker results by Demartini et al. (2013).

We compute the posterior as the maximum conditional probability of the incentive v at a given point x using Bayesian inferencing as shown in the equation below:

$$\Pr(v|x) = \frac{\Pr(x|v)\Pr(v)}{\Pr(x|v)\Pr(v) + \Pr(x|\neg v)\Pr(\neg v)} \tag{8.1}$$

$$\Pr(v|x, s = in) = \frac{\Pr(x|v, s = in)\Pr(v|s = in)}{\Pr(x|s = in)} \tag{8.2}$$

where:
v is a potential incentive to be shown to the player.
x is the number of images a player has tagged so far.
s is the state of a player being in or out of the game.
Pr(v—x,s=in) is the posterior of the incentive given the number of images the player has tagged.
Pr(v—s=in) is the prior probability of the incentive i.e., the probability of any player at any given point accepting this incentive.
Pr(x—v,s=in) is the likelihood at the current game point that the player would accept the given incentive.

8.2.5.1 Definitions

Image point
$x \in X = \{1, ..., N\} where N = 2,200.$
Represents the number of images a player has tagged at the point of an attempted game exit.

Game states
$s \in S = \{out, in\}$
Represents the state a game player is in after attempting to exit the game at an image point.

Incentive
$v \in V = \{badges, leaderboard, levels, access, power, money\}$
Represents the set of incentives from which v is drawn to be presented to the player at the point of attempted exit.

Image Band
$b = (x_i, x_j) = \{b \in \mathbb{R} \mid x_i < b < x_j\}$

Represents a range over image points x_i to x_j over which players exhibit similar exit pattern behaviours.

8.2.5.2 Prior Distributions

Our prior distributions come from the results of random incentives presented to players. We compute the *objective prior* of the incentive v as given by the sum and product rule in Bayes Theorem:

$$\Pr(v) = \sum_{x=1}^{N} \Pr(v, x) \tag{8.3}$$

$$\Pr(v, x) = \Pr(v) \Pr(x|v) = \Pr(x) \Pr(v|x) \tag{8.4}$$

$$\Pr(v|s = in) = \frac{\sum\limits_{x=1}^{N} \Pr(s = in|v, x)}{\sum\limits_{x=1}^{N} \sum\limits_{v \in V}^{V} \Pr(s = in|v, x)} \tag{8.5}$$

This represents the number of players that remained in state $s = in$, at image point x after being shown incentive v over all image points x ∈ X, compared with all the players that remained in state $s = in$, at image point x after being shown any incentive v in the set of all incentives V over all image points x ∈ X.

As an example, given that 100 players remained in state $s = in$ after they were shown any incentive v ∈ V over all image points x ∈ X = {1, ..., 2,200}. If 29 of such players (who remained in the game) were shown incentive v = '*power*', then the prior of the incentive '*power*' over all image points is 29/100 or 29%.

8.2.5.3 Likelihood Distributions

We also compute the likelihood at each image point x, of a worker remaining in state $s = in$ given an incentive v. The likelihood represents the conditional probability

$$P(x|v, s = in)$$

at image point x. Our likelihood function was a product of 2 variables: (a) the image point likelihood and (b) the image band likelihood.

Image Point Likelihood

For each incentive v, we calculate the image point likelihood at point x as:

$$\Pr(x|v, s = in) = \frac{\Pr(s = in, x|v)}{\sum\limits_{s \in S}^{S} \Pr(s, x|v)} \tag{8.6}$$

This represents how many players remained in state $s = in$, at image point x after being shown incentive v at image point x, compared with all observations of state changes at image point x after being shown incentive v.

As an example, given that 3 players attempted to stop playing the game after tagging 11 images (image point x = 11) and the 3 players were all shown the incentive v = '*power*'. If 2 of the players go on to tag the 12th image, then we calculate the likelihood of a player remaining in state s=in, at the 11th image, when shown 'power' as 2/3.

For image points where we do not have any observed behaviours, i.e., where no player had attempted to exit the game at a certain image point x, we apply the principle of indifference (*principle of maximum entropy*) to accommodate these latent variables.

$$P(x|v, s = in) = 1/N \;\forall\; x \in N = \{1, ..., N\} \tag{8.7}$$

The variable x here represents the image point while N = 2, representing 2 possible states s = {in, out}. Therefore, for unobserved image points,

$$P(x|v, s = in) = P(s = in) = P(s = out) = 0.5 \tag{8.8}$$

Image Band Likelihoods

To further accommodate for latent variables and present an expressive picture of how players behave after tagging a certain numbered range of images, we introduced image band likelihoods.

Image Band

$b = (x_i, x_j) = \{b \in \mathbb{R} \mid x_i < b < x_j\}$

Represents a range over image points x_i to x_j over which players exhibit similar exit pattern behaviours.

The image bands b ∈ B were elicited by observations over the results from the randomised incentive condition and they are:

$$B = \{0 - 11, 12 - 60, 61 - 100, 101 - 200, 201 - 2200\}$$

The image band likelihoods were computed on an incentives basis, as such:

$$\sum_{b \in B}^{B} \Pr(s = in, b|v) = 1 \tag{8.9}$$

For each incentive v, we calculate the image band likelihood over band b as:

$$\Pr(b|v, s = in) = \frac{\Pr(s = in, b|v)}{\sum\limits_{s \in S}^{S} \Pr(s, b|v)} \tag{8.10}$$

This represents how many players remained in state s = *in*, within image band b, after being shown incentive v, compared with all players who remained in state s = *in*, over all image bands b ∈ B, after being shown incentive v.

As an example, given that 100 players remained in state s = *in* after they were shown incentive v = '*power*' over all image points x ∈ X = 1, ..., 2,200. If 16 players go on to tag 1 more image within the range of image points x_i, x_j = (12,60) = 12 ¡ b ¡ 60, then the image band likelihood of '12-60' given incentive v = '*power*' is 16/100 or 16%.

8.2.5.4 Updating the Likelihoods

As the experiment runs, we continuously take into account the behaviour of players at each image point. With each new observation at an image point, we recalculate the likelihood of remaining in state s = *in*, at image point x after being shown incentive v. Therefore, our probabilistic model iteratively updates the likelihoods by constantly learning and taking into account new data based on player interaction. This is of particular importance in filling in revised parameters for the earlier unobserved image points.

As an example, given an image point x = *20*, where there had been no earlier observations in *experiment 4* of a player exit after being shown incentive v = '*power*'. The image point likelihood would be assigned the default of 0.5 (*principle of maximum entropy*), computed as 2 observations with 1 observation at state s = *in*. If a new observation occurs (for any given player) at the image point x = *20* for incentive v = '*power*' and the player transitions to state s = *out*, the image point likelihood is updated to 3 observations with 1 observation at state s = *in* = 0.33.

8.2.5.5 Computing the Posteriors

Given the incentive prior P(v—s=in) when a player remains in the game, the image point likelihood $\Pr(x|v, s = in)$ and the image band likelihood $\Pr(b|v, s = in)$, we are able to compute the best incentive to offer a player at image point x as the incentive that maximizes the posterior given as

$$\arg\max_{v} \Pr(v|x) = \Pr(j|v, s = in) \Pr(v|s = in) \tag{8.11}$$

where the joint likelihood of the image point and the image band is given as:

$$\Pr(j|v, s = in) = \Pr(x|v, s = in) \Pr(b|v, s = in)$$

The incentive is then offered to the player and the ensuing state transition is recorded as a new observation point to update the image point likelihood given that incentive.

8.2.5.6 Algorithms

We now present the algorithm for the image point likelihoods for an incentive and the algorithm for calculating the incentive posteriors at any given image point.

```
Result: Likelihood P(x—v) = inx/obv
Parameter: v = incentive;
Initialize Latent Variables;
Image Points: x = {1,...,N};
Observations at x: obv = 2;
State = in at x: inx = 1;
for x in Image Points do
    if state = in at x then
        obv += 1;
        inx += 1;
    else
        obv += 1;
    end
end
```

Algorithm 2: Image point likelihoods for incentive v

```
Result: Posterior Incentive: $\arg\max_v \Pr(v|x)$
Parameter: x = image point;
Initialize Latent Variables;
Incentive v at x: vx = {};
Posteriors Tracker: pt = {levels = 0, ..., power = 0};
Min Tags: min = 11;
Max Tags: max = 2,200;
for image tag x from min to max do
    $\Pr(v|x)\ \forall\ v \in V$;
    $\Pr(v|x) = \Pr(v).\Pr(b|v).\Pr(x|v)$;
    Incentive Identifier iid = 0;
    Selected Incentive vx = $\Pr(v|x)$ at iid;
    Update Posteriors Tracker $ptatvx+ = 1$;
    Max Incentive Assignment $mia = \Pr(v) * (max - min)$ if pt at vx ¡ mia then
        return vx;
    else
        return vx = $\Pr(v|x)$ at iid + 1;
    end
end
```

Algorithm 3: Incentive posteriors at image point x

8.3 Experiment Design

This section details the experiments we carried out. We ran 2 experiments. Experiment 1 had 2 conditions - (a) CrowdFlower (non gamified) condition; (b) Wordsmith (gamified) condition. Experiment 2 had 4 conditions - (a) Non-gamified condition; (b) Gamified condition (without furtherance incentives); (c) Gamified condition (with random furtherance incentives); (d) Gamified condition (with targeted furtherance incentives).

8.3.1 Research Questions

Our work was centered around 3 potential ways in which gamification can be used to improve paid microtasks:

1. Gamifying paid micro-tasks leads to increased worker engagement, culminating in more work done for less cost.
2. Gamifying paid micro-tasks leads to higher inter-annotator agreement, yielding higher quality results than without.
3. Targeting incentives when a player attempts to quit the task leads to increased engagement.

To test these hypothesis, we carried out 2 experiments in image labelling. Workers were presented with an image, and asked to assign keywords that describe the image. To test the first two hypotheses, we chose a between-subjects design where the control condition consisted of a standard, non-gamified interface, using CrowdFlower's image labelling job; while the experimental condition, consisted of a gamified interface incorporating several game elements. Both conditions relied on CrowdFlower for worker recruitment, but while workers performed tasks directly within CrowdFlower for the control condition, workers assigned the experimental were redirected to an external game site. Participants in both setups were paid the same amount.

To test the 3rd hypothesis, we carried out 2 additional condition setups on our gamified interface, again with players sourced and redirected from CrowdFlower. In the control condition, workers were presented with a randomly-selected incentive to stay when they attempted to leave the game. In the experimental condition, an incentive was shown, selected based upon a predictive model constructed from the previous worker's task history. The details of this predictive model was described earlier in Section 8.2.5.

8.3.2 Research Data

For our experiments, we used the ESP game dataset from von Ahn and Dabbish (2004). This comprises of 100,000 images and about 1.4 million image tags. For each experiment, we selected the images in the dataset which had the highest number of keyword tags associated with it. This was used as a sort of quasi-ground truth, for checking basic tagging quality and assigning bonus points.

8.3.3 Evaluation Metrics

To evaluate worker performance, we measured both the volume of work completed and work quality. To assess the volume of work completed, we simply measured the average number of keywords provided per image. To assess work quality, we used two measures: overlap with the gold standard keywords in the dataset, and a standard measure of inter-annotator agreement from Bhowmick et al. (2008) to determine the degree of the pairwise consensus of image labels which were not in the gold standard datset.

We use the approach by o determine the pair-wise agreement. Given **I** as the number of images, **K** is the total number of annotations for an image, **H** is the number of human agents (crowd workers or game players) that annotated the image and **S** is the set of all keyword pairs with cardinality $|S| = \binom{K}{2}$, where $k_1 = k_2 \ \forall \ \{k_1, k_2\} \in S$.

Given an image $\boldsymbol{i}$ and an assigned keyword $\boldsymbol{k}$ where $\{k, k\} \in S$, the average agreement, A_{ik}, on the keyword $\boldsymbol{k}$ for the image $\boldsymbol{i}$ is given by

$$A_{ik} = \frac{n_{ik}}{\binom{H}{2}} \tag{8.12}$$

where n_{ik} is the number of human agent pairs that agree that keyword $\boldsymbol{k}$ describes image $\boldsymbol{i}$.

Therefore, for a given image $\boldsymbol{i}$ the average agreement over all assigned keywords is

$$A_i = \frac{1}{|S|\binom{H}{2}} \sum_{k \in S}^{S} n_{ik} \tag{8.13}$$

For Experiment 2 condition 4, we sought to evaluate the effectiveness of targeted incentives over random ones. To do this, we used as a measure the number of players that tagged at least 1 more image after they were presented with each particular incentive. The incentive was shown when they attempted to stop playing the game.

8.3.4 Experimental Conditions

In this section, we summarise both experiments and their conditions in detail.

Experiment 1: *Task: Tag 1 Image with at least 2 keywords; Source dataset size: 200 images; Workers: 600; Payment: \$0.02; Platforms: CrowdFlower and Wordsmith.* In the first experiment, workers in either condition were required to tag 1 image with 2 keywords. In Wordsmith, the gamified condition, this corresponded to advancing 1 level into the game. Players in Wordsmith could continue playing the game (tagging more images) after completing the required annotation. There were 200 images in the dataset. Participants were paid 2 cents for the image tagged.

Experiment 2: *Task: Tag 11 images with at least 2 images each; Source dataset size: 2,200; Workers: 600; Payment: $0.10; Platforms: Crowdflower and Wordsmith; Furtherance Incentives:none, random or targeted.* In experiment 2, workers were required to tag 11 images with keywords. However, the dataset size was increased 11 fold (from 200 to 2,200) to allow players to play for longer without seeing repeated images. Intermediate results had revealed that a certain number of players tagged the entire dataset of 200 images. This experiment consisted of 4 conditions detailed below. In addition, for conditions 3 and 4, furtherance incentives, defined in Section 8.2.4 are introduced when players attempt to quit.

Experiment 2 - Condition 1: *Platform: CrowdFlower; Furtherance Incentives: none.* This was a non-gamified setup where workers were required to tag 11 images from a dataset of 2,200 images for 10 cents.

Experiment 2 - Condition 2: *Platform: Wordsmith; Furtherance Incentives: none.* In this gamified setup, players were required to tag 11 images from a dataset of 2,200 images for 10 cents. This advanced them 2 levels into the game. The players could continue tagging (playing the game) if they wished.

Experiment 2 - Condition 3: *Platform: Wordsmith; Furtherance Incentives: Random.* Identical to Condition 2, except a random furtherance incentive is presented when a player attempted to exit the game.

Experiment 2 - Condition 4: *Platform: Wordsmith; Furtherance Incentives: Targeted.* Identical to Condition 3, except that the furtherance incentive was selected according to the maximum likelihood of user retention using the probabilistic model presented in Section 8.2.5.

8.4 Results

The result of Experiment 1 is summarised in Table 8.2. The results illustrate that players (participants in the game condition) supplied more keywords than those in the CrowdFlower condition on average (97 per image vs. 2), and labelling more images overall (32 per worker vs. 1), resulting in an overall yield of 41,206 total keywords vs. 1,200 in the control condition. We note that, since the control condition restricted workers to supply up to two keywords for a single image, it is unsurprising that individuals in the control condition provided only two keywords for a single image. However, in both conditions, individuals were rewarded only up through the same amount of work (completing the task of supplying 2 keywords for a single image), and thus the additional work done in the Wordsmith condition was not financially incentivised and done for free. Moreover, compared to the control, the experimental condition yielded significantly more *new keywords*, which we define to be keywords that were not in the original gold standard seed, but achieved the requisite threshold of inter-annotator agreement. The average inter-annotator agreement, computed as described in 8.3.3 over all images for the control condition was also much less than that of the experimental condition (5.72% vs. 37.7%). The control condition achieved 42.9% coverage of the original gold standard label set, while the experimental condition covered 52.5%.

Table 8.3 summarises the results for Experiment 2. Again, compared with the CrowdFlower interface, all game conditions saw much greater output, both in terms of labels per image (average 40,510 keywords across game conditions vs. 13,200 in the control condition) and number of images tagged (30 images labeled per worker across game conditions vs. 11) despite monetary compensation being held constant between conditions (10 cents to complete 11 images with 2 labels each). Examining the game conditions only, conditions 3 and 4 which featured furtherance incentives on exit attempt resulted in players performing more labels on average (31.5 vs. 27) than condition 2, which had no furtherance incentives. We note that due to the much larger source dataset of images, the likelihood that two workers would be presented the same image is much lower, resulting overall in noisier inter-rater agreement and lower coverage of gold-standard labels.

To analyse player response to furtherance incentives, Table 8.6 lists the number of players who responded to each furtherance incentive stimulus type at various levels of play (image bands). To clarify, we considered a player to be *responding to the incentive stimulus* when, upon attempting to quit the game *and* being presented with a furtherance incentive, decided to tag at least 1 extra image prior to exiting. The table indicates the number of responses of the number of presentations of each stimuli for each (C3 random and C4 targeted) condition at 5 image image bands, corresponding to the number of images previously tagged when attempting to exit. Comparing randomised to targeted incentive, the results point to greater response to furtherance incentives when delivered in the targeted incentive condition (C4) than randomised (C3). In the targeted incentives condition, 77% of players went on to tag at least 1 more image, compared with only 27% in the randomised condition.

With respect to furtherance incentive type, direct comparison is in Table 8.6 due to the fact that the number of stimulus presentations differ for different types and conditions. We constructed Table 8.4 to make this comparison further, which simply presents a breakdown, by type, of all successful furtherance incentive stimulus responses. As can be seen, in both C3 (Randomised) and C4 (Targeted), the Power and Money incentives made up the top two successful incentives, with Money comprising the largest share of the targeted successes. We discuss these results in the next section.

Experiment 1		
Metric	CrowdFlower (control)	Wordsmith (experimental)
Total workers	600	423
Total keywords	1,200	41,206
New keywords	111	5,708
Avg. agreement	5.72%	37.7%
Gold keywords	42.92%	52.53%
Mean Imgs/person	1	32
Max Imgs/person	1	200

Table 8.2: *Experiment 1 Results* - High level results for Experiment 1, comparing number of keywords and images tagged in the gamified (Wordsmith) condition compared to the standard CrowdFlower interface.

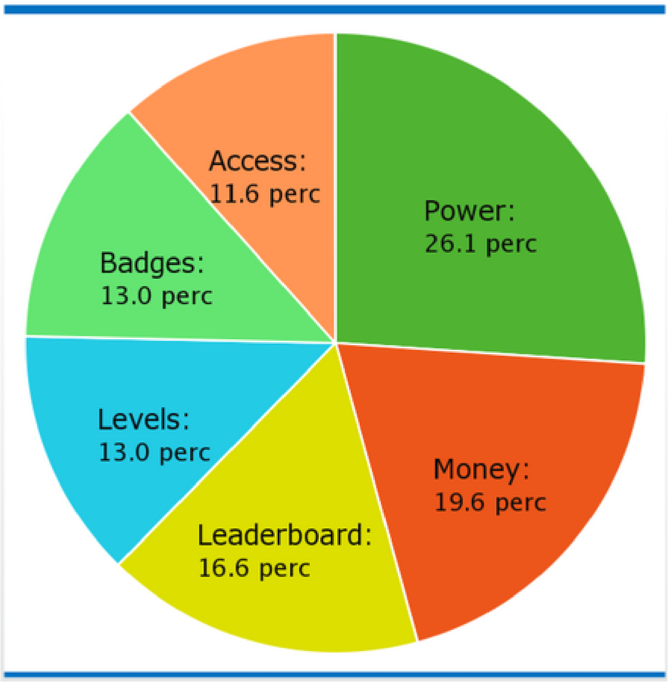
Experiment 2				
	CrowdFlower	Wordsmith (Gamified)		
	C1: No furtherance	C2: No furtherance	C3: Random furtherance	C4: Targeted furtherance
Total workers	600	514	543	454
Total keywords	13,200	35,890	47,418	38,223
New keywords	1,323	4,091	5,435	4,849
Avg. agreement	6.32%	10.90%	10.16%	9.86%
Gold keywords	48.42%	45.02%	41.21%	47.10%
Mean Imgs/person	11	27	33	30
Max. Imgs/person	11	351	501	540

Table 8.3: *Experiment 2 Results* - High level summary of work output and quality comparing non-gamified (C1) and gamified (C2, C3, C4) conditions.

Incentive	C3: Randomised	C4: Targeted
Power	26.09%	30.16%
Money	19.65%	46.17%
Leaderboard	16.59%	5.71%
Levels	13.01%	7.34%
Badges	13.04%	5.98%
Access	11.61%	4.35%

Table 8.4: *Incentive Response Distribution* - Successful furtherance incentives stimuli broken down by type, for both C3 (randomised) and C4 (targeted) conditions.

Figure 8.3: Incentive Type Distribution in Random Furtherance Incentives Condition



Condition	Uptake
Random Incentives	27.43%
Targeted Incentives	76.55%

Table 8.5: Incentives Uptake

Image band	11		12-60		61-100		101-200		201-2,200	
	Rand	Target	Rand	Target	Rand	Target	Rand	Target	Rand	Target
Power	26.67% (4/15)	70.97% (22/31)	33.33% (9/27)	74.68% (59/79)	55.55% (5/9)	80.95% (17/21)	25.00% (2/8)	100.00% (5/5)	100.00% (3/3)	61.54% (8/13)
Money	23.53% (4/17)	88.24% (75/85)	34.78% (8/23)	77.57% (83/107)	66.67% (2/3)	100.00% (2/2)	40.00% (2/5)	100.00% (6/6)	0.00% (0/0)	100.00% (5/5)
Leaderboard	21.43% (3/14)	0.00% (0/1)	10.00% (3/30)	70.00% (7/10)	33.33% (1/3)	100.00% (1/1)	70.00% (7/10)	57.14% (12/21)	75.00% (3/4)	100.00% (1/1)
Levels	13.04% (3/23)	40.00% (2/5)	18.18% (6/33)	69.57% (16/23)	75.00% (3/4)	100.00% (3/3)	16.67% (1/6)	100.00% (3/3)	100.00% (2/2)	100.00% (3/3)
Badges	0.00% (0/0)	0.00% (0/1)	22.22% (4/18)	63.63% (7/11)	16.67% (1/6)	100.00% (3/3)	66.67% (4/6)	90% (9/10)	66.67% (2/3)	75.00% (3/4)
Access	11.76% (2/17)	0.00% (0/0)	17.24% (5/29)	33.33% (5/15)	33.33% (3/9)	50.00% (2/4)	20.00% (1/5)	100.00% (3/3)	100.00% (1/1)	100.00% (6/6)

TABLE 8.6: *Furtherance Incentive responses (Results of Experiment 2 Condition 3 & 4)* Percentage of players in who responded to each Furtherance Incentive broken down by type (Power, Money, Leaderboard, Levels, Badges and Access), and condition (randomised vs. targeted). The number of incentive prompts delivered for each type are listed by the responded percentage.

Metrics	Access	Badges	Leaderboard	Levels	Money	Power
Total results						
Observations	61	49	61	68	48	62
State=in at x	12	11	17	15	16	23
Band 11						
Observations	17	0	14	23	17	15
State=in at x	2	0	3	3	4	4
Band 12 - 60						
Observations	29	18	30	33	23	27
State=in at x	5	4	3	6	8	9
Band 61 - 100						
Observations	9	6	3	4	3	9
State=in at x	3	1	1	3	2	5
Band 101 - 200						
Observations	5	6	10	6	5	8
State=in at x	1	4	7	1	2	2
Band 201 - 2200						
Observations	1	3	4	2	0	3
State=in at x	1	2	3	2	0	3

Table 8.7: Incentives Distribution in Image Bands based on the results of the randomised furtherance incentives condition (Experiment 2 Condition 3)

8.5 Discussion

In this section, we first briefly re-visit our results in the context of the research hypotheses, discussing limitations in the process. We then discuss implications of our findings to crowdsourcing, and conclude with a summary of ongoing and future work.

The results demonstrate support for all three of our research hypotheses. With respect to H1, players in the game condition unilaterally performed more tasks even when they were not explicitly incentivised with monetary reward to do so. In addition, output was of higher quality, indicating support for H2, both when measured in terms of diversity (new words with high agreement) and achieved consistent coverage of the gold standard labels than the control condition. In particular, we saw no support for overjustification in these results, which would have been manifest in reduced productivity with the introduction of game elements.

One limitation of our experimental design is that, since the number of contributions in the control interface was clamped while the game condition was not (meaning they could contribute indefinitely), it is not meaningful to quantify the increased volume of work between the control and gamification conditions. However, we can compare quality differences (which signalled significant gains), and volume differences among just the game conditions in Experiment 2, when targeted furtherance incentives were shown to yield higher volumes of work than randomised ones (H3).

However, perhaps more significantly, this study demonstrated that even simple furtherance incentivisation methods do work towards getting players to complete more tasks. In all but the Money furtherance incentive condition, such methods worked to increase output at no extra cost. Moreover, we found that among furtherance incentivisation strategies, those that were more social generally fared better than those that were personal; for example, the Power incentive was presented '*You would be rewarded with the power to view other players tags*', while the Leaderboard incentive promised participants a higher place on the leaderboard, which was visible to everyone. This agrees with previous work in GWAPs such as the ESP Game, in particular von Ahn and Dabbish (2004)) in which social incentives were evidenced to be among the most powerful. Most human computation environments, like Mechanical Turk, CrowdFlower and citizen science projects still lack elements that promote social visibility that might improve engagement.

The effectiveness of money as an effective furtherance incentive was somewhat surprising, given the fact that most participants already performed free labour, that is, work beyond the minimum that was asked of them to get their initial reward. Therefore, it could be concluded that these participants were motivated to do this additional work for other reasons. However, when financial reward is re-introduced as a furtherance incentive, it effectively motivated people to complete more work. Further analysis is required to understand to what extent such monetary rewards could compel continued participation, and the optimal amounts of reward for doing so.

As our experiments only tested one type of crowdsourced task and GWAP, namely image labelling, the results may not necessarily apply to all task types. In particular, tasks that require high cognitive load, require significant time investment or creative thought may not benefit from game mechanics due to their intensive nature. Moreover, those kinds of crowdsourced tasks driven by strong intrinsic motivations (such as citizen science, disaster relief, and so on) are unlikely to substantially benefit from these results because such motivations will probably overshadow the simpler incentives tested here. Moreover, such intrinsic motivation settings have been shown to be more prone to overjustification effects, and thus may result in actually reduced participation. We wish to test whether such effects will become present in such settings in future experiments.

Among our ongoing efforts, we wish to better understand how and why the incentives work in the ways and to the extent that they do. In particular, we believe that furtherance incentives could be more effective if carefully distributed within the game mechanics so that they appear at appropriate intervals when motivation begins to wane, not only after the participant has initiated an attempt to leave.

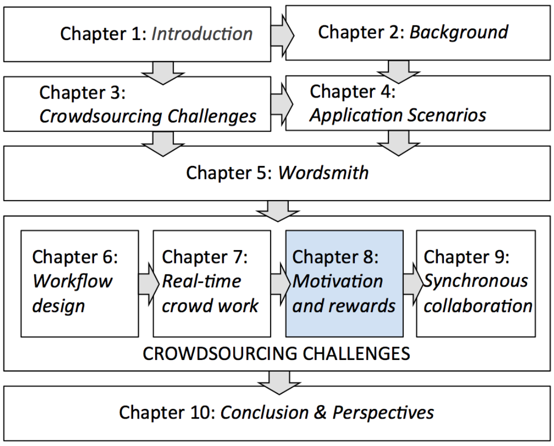
Second, we wish to the improve the probabilistic model to take into account other aspects of players' performance, task history and demographic, and to understand the ways in estimating appropriate rewards. In particular, we wish to run further experiments to determine whether incentives are more effective for particular demographics than for others, or for workers at particular skill levels or task completion histories.

In addition, we would like to investigate further social effects of furtherance incentives. In this experiment, levels and badges were merely to mark a player's own progress; however, these might be made more effective if such rewards were made visible to other players and seen as a form of status. Such status has been shown to effectively encourage participation in online communities (Kraut et al., 2012) and may translate well to micro-task environments as well. Moreover, all of the incentives we applied in this experiment were positive, individual incentives; we next wish to explore the effectiveness of other types of incentives such as positive social incentives (e.g., members of an entire group get a reward), as well as negative incentives both as in-game elements and furtherance incentives. Finally, we would like to understand the span of furtherance incentives, and potential avenues for extending the effects of such incentives in various ways.

8.6 Conclusion

In summary, our results have illustrated that by adding gamification to crowdsourced micro-tasks that already have external incentives, we can improve the quality and quantity of work completed. Our results complement previous work comparing purely gamified and paid crowdsourcing, and extend previous results with a look at multiple kinds of furtherance incentives, combined with reward adaptation. Although we have shown social incentives and supplemental monetary rewards outperform other such incentives, and demonstrated a simple probabilistic model able to outperform a randomised strategy, we believe that we have only begun to understand the relationships that such incentives have on subjective worker experience and sustained engagement in the long term, and plan to continue to pursue such investigations in the future.

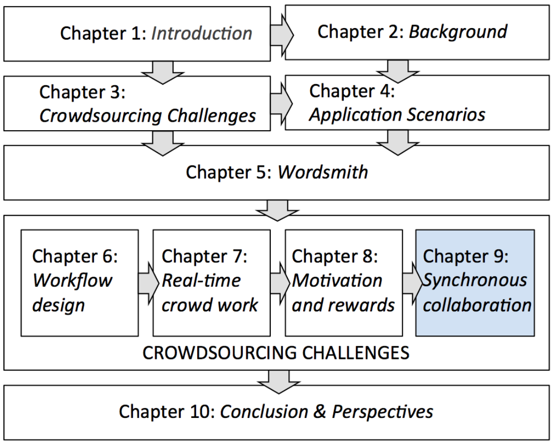
8.7 Summary


Chapter 1: *Introduction*
Chapter 2: *Background*
Chapter 3: *Crowdsourcing Challenges*
Chapter 4: *Application Scenarios*
Chapter 5: *Wordsmith*
Chapter 6: *Workflow design*
Chapter 7: *Real-time crowd work*
Chapter 8: *Motivation and rewards*
Chapter 9: *Synchronous collaboration*
CROWDSOURCING CHALLENGES
Chapter 10: *Conclusion & Perspectives*

In this previous chapter, we discussed the challenge of motivation and rewards in gamified paid microtask crowdsourcing. The chapter studied the effects of different gamification mechanisms layered on micro payments as a tool to improve worker engagement and increase task quality. The chapter also completed our understanding of furtherance incentives by demonstrating how to combine the two components that make furtherance incentives possible i.e., predicting worker exit and applying appropriate incentives.

Chapter 9

Synchronous Collaboration


Chapter 1: *Introduction*
Chapter 2: *Background*
Chapter 3: *Crowdsourcing Challenges*
Chapter 4: *Application Scenarios*
Chapter 5: *Wordsmith*
Chapter 6: *Workflow design*
Chapter 7: *Real-time crowd work*
Chapter 8: *Motivation and rewards*
Chapter 9: *Synchronous collaboration*
CROWDSOURCING CHALLENGES
Chapter 10: *Conclusion & Perspectives*

In this chapter, we address the fourth challenge of 'synchronous collaboration'. In this chapter, we also apply the concept of furtherance incentives in continuance with the insights gained from the previous chapters. In particular, in the last chapter (Chapter 8), we observed that sociality based incentives were the most effective drivers of retention and engagement. We therefore expand our knowledge by experimenting with two sociality-driven furtherance incentives – social pressure and social flow in our study of synchronous collaboration in microtask crowdsourcing.

This chapter is adapted from earlier published work [1] titled 'Please Stay vs. Lets Play: Social Pressure Incentives in Paid Collaborative Crowdsourcing'.

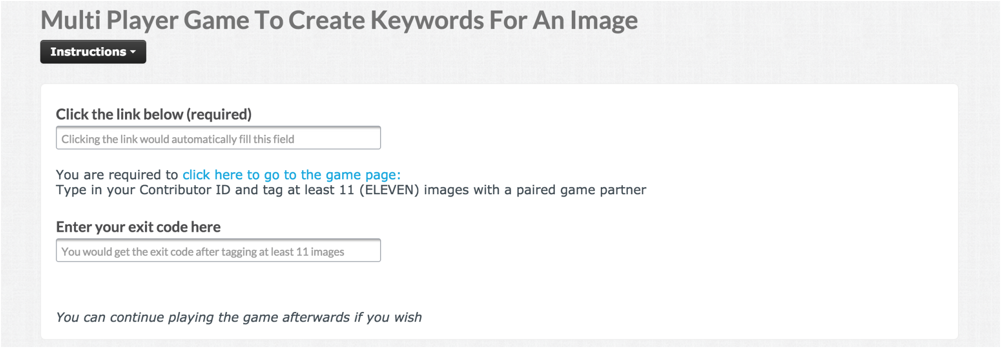
9.1 Overview

In this chapter we continue our discussion by experimenting with a new model of 'groupsourcing' (Chamberlain, 2014a), which relies on financial rewards paired with social flow and social pressure as drivers for more accurate answers, improved engagement, and cost savings. In the experiments we use Wordsmith which supports both a 'traditional' (single-worker) and a collaborative microtasking model. In the latter, each task is executed synchronously by a pair of workers, who have to agree on the solution without

[1] This chapter is adapted from work that appeared at ICWE 2016 Feyisetan and Simperl (2016)

communicating with each other. Rewards are granted individually; however, task completion, and hence being paid, depends on the willingness of peer workers to continue to engage. Workers receive a payment as long as they complete a minimum number of tasks. In the same time, they are free to leave before reaching their pre-defined targets or continue annotating after achieving them. Workers are recruited from existing paid microtask platforms; the experiments reported in this chapter are using CrowdFlower.

Our basic hypothesis was that Wordsmith's collaborative model is attractive for a community of image labelling workers who are used to solving tasks on their own and being rewarded based on individual performance. To test this hypothesis we ran experiments in which we compared the behaviour of the crowd when carrying out tasks the traditional (non-collaborative) way or in pairs seeking consensus. The results were extremely encouraging. The collaborative condition attracted a significant number of CrowdFlower workers - we put out a call for 600 workers and received over 2000, which were paired into more than 9000 teams that labeled 4500 images, 233% more than the targets we asked for. We worked with two variations of image labelling tasks, in which we increased the complexity of the task (low threshold, LT: one image vs. high threshold, HT: 11 images) and adjusted the prices accordingly, observing a similar trend. In addition, we studied whether a task design which enforces consensual answers will increase task accuracy and output within the same budget. The experiments revealed significant increase in the latter (measured by the average number of tags annotated and the average number of new tags generated) by microtask workers engaging collaboratively than individually. This result was consistent in the two task conditions ($LT = 132$ *vs* 54 tags per person; $HT = 218$ *vs* 54). Collaborative workers also recorded higher inter-annotator agreement scores ($LT = 35\%$; $HT = 26\%$) than those annotating alone ($LT = 29\%$; $HT = 14\%$).



Multi Player Game To Create Keywords For An Image

Click the link below (required)

Clicking the link would automatically fill this field

You are required to click here to go to the game page:
Type in your Contributor ID and tag at least 11 (ELEVEN) images with a paired game partner

Enter your exit code here

You would get the exit code after tagging at least 11 images

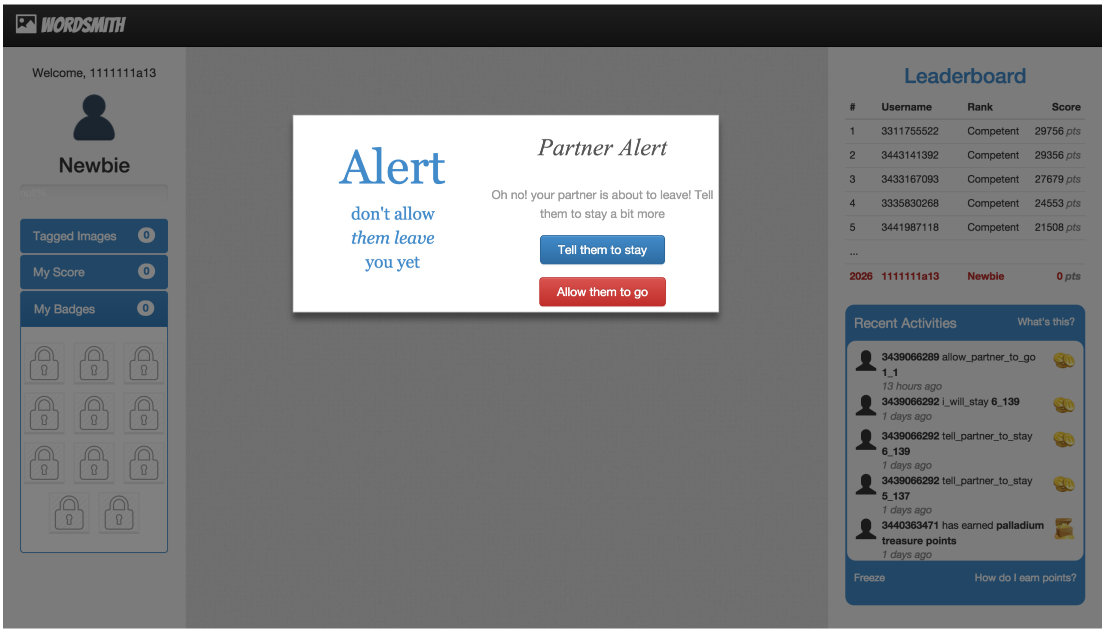
You can continue playing the game afterwards if you wish

FIGURE 9.1: CrowdFlower Request Interface

Finally, we went on and looked at the effects of socially motivated incentives on workers behaviour. To do so, we ran a new round of experiments in which a worker who is about to leave can be prompted by their co-worker to stay and continue to tag images. This might be helpful, among other things, if one of the collaborators has not yet reached their target in terms of number of tasks completed, and risks missing the reward; or if the co-tagging experience is so entertaining that people would prefer to enjoy it a bit longer. Our results indicate the importance of these social incentives, particularly empathic social pressure, wherein a worker continues the task, despite receiving their own payment, in order to help their partner complete enough task to receive their own payment. We also recorded an increase in task output (312 vs. 218 tags per person) and inter-annotator agreement (29% vs. 26%) when these incentives were present compared to when they were not.

9.2 Model

In the collaborative mode of Wordsmith, a worker's advancement is tied to the cooperation and effort of their partner. At the first launch of the task, the initial cohort of workers join at the same level: newbies with zero points; hence, workers are incentivised to help each advance to the next levels which result in individual payoffs. This is in line with incentives generated by people having shared circumstances (i.e., the need to get paid), as stated by Kandel and Lazear (1992). However, as existing workers transition to higher levels, and partner switching becomes more frequent, they (the existing workers who have almost completed their task) might be oblivious or less inclined to help new workers advance through the task.



Oh no! your partner is about to leave! Tell them to stay a bit more

#	Username	Rank	Score
1	3311755522	Competent	29756 *pts*
2	3443141392	Competent	29356 *pts*
3	3433167093	Competent	27679 *pts*
4	3335830268	Competent	24553 *pts*
5	3441987118	Competent	21508 *pts*
...			
2026	1111111a13	Newbie	0 *pts*

3439066289 allow_partner_to_go **1_1**
13 hours ago

3439066292 i_will_stay **6_139**
1 days ago

3439066292 tell_partner_to_stay **6_139**
1 days ago

3439066292 tell_partner_to_stay **5_137**
1 days ago

3440363471 has earned **palladium treasure points**
1 days ago

FIGURE 9.2: Wordsmith Partner Alert

When the *social incentive* setting is activated in the collaborative mode in Wordsmith, a worker is given a heads up when their partner is about to quit the task. The worker can then select one of two options: *tell them to stay* or *allow them to go*. Choosing to request their partner to remain in the task represents the cost of exerting the social pressure as presented by Calvó-Armengol and Jackson (2010). The message which is then automatically conveyed to the partner is dependent on the sending worker's current level in the task: the message either appeals to the partner to continue till the requesting worker reaches the level where they receive a payoff (*please stay*) or, the message requests that both workers continue annotating for fun if the payoff has been received by the requesting worker (*let's play*).

9.2.1 Task

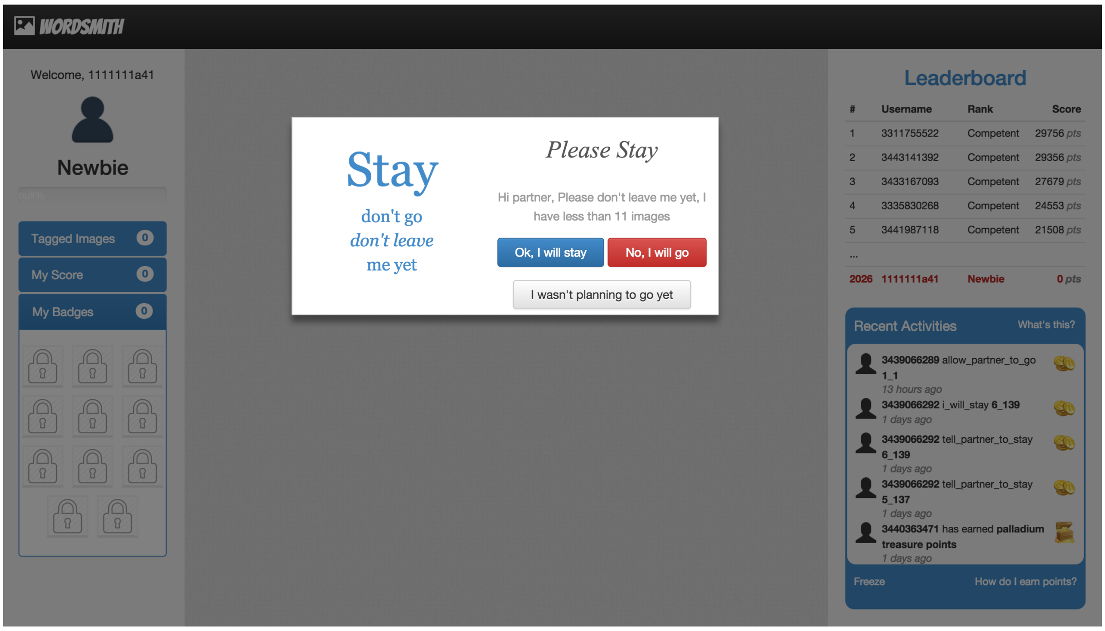
The participants were required to annotate a given image with a set of descriptive keywords. In the traditional single worker mode, workers simply needed to input a set of valid labels, while in the collaborative mode, workers were paired, and required to correctly guess and match on a set of keywords. We selected this task because the task domain had been well studied in literature, and we had access to a sizable dataset.

9.2.2 'Please Stay': The Role of Social Pressure as an Incentive

There are two broad possible scenarios in which a worker sends a *please stay* message to their partner: the worker is at the same level with their partner (shared circumstances in which neither worker has been paid for completing the task), or, the worker is at a lower level than their partner (in the scenario where the partner has been paid and is just annotating for fun). In either case, the immediate continuity of the requesting worker, and their potential to get paid, is dependent on their partner deciding to remain in the task. Kandel and Lazear (1992) stated that, *incentives are generated when an individual empathizes with those whose income he affects*. In addition, according to Eisenberg and Miller (1987), empathy is an affective state the stems from the comprehension of another's emotional condition, which is in harmony with it (i.e., my emotional state is congruent with yours). They demonstrated empirically that empathy could have positive associations with pro-social behaviour.

The partner receiving the request: 'Please stay, I have not yet tagged x images, Don't leave me yet'; is then presented with a set of options: '*OK, I will stay*', or '*NO, I will go*'. The subsequent actions of the receiving partner are not globally observable, and this underscores the interplay and difference between *guilt* and *shame* (Tangney and Dearing, 2003) as sources of social pressure. For example, a partner who has been paid, and selects '*NO, I will go*', leaving the original worker without a partner, feels no external shame since the action was not observable by all other workers. Without observability, only the internal guilt of leaving a fellow worker unpaid, or empathy stemming from recently being in the same situation, serve as a form of social pressure.

A receiving partner that chooses to opt out of the task before receiving their own reward might feel less guilt in leaving the requesting worker hanging, afterall, they haven't been



Stay

don't go
don't leave
me yet

Please Stay

Hi partner, Please don't leave me yet, I have less than 11 images

Leaderboard

#	Username	Rank	Score
1	3311755522	Competent	29756 *pts*
2	3443141392	Competent	29356 *pts*
3	3433167093	Competent	27679 *pts*
4	3335830268	Competent	24553 *pts*
5	3441987118	Competent	21508 *pts*
...			
2026	**1111111a41**	**Newbie**	**0** *pts*

Recent Activities

3439066289 allow_partner_to_go **1_1**
13 hours ago

3439066292 i_will_stay **6_139**
1 days ago

3439066292 tell_partner_to_stay **6_139**
1 days ago

3439066292 tell_partner_to_stay **5_137**
1 days ago

3440363471 has earned **palladium treasure points**
1 days ago

FIGURE 9.3: Wordsmith Please Stay Alert

paid also. Furthermore, according to Kandel and Lazear (1992), *guilt may require a greater amount of past investment than shame.* A worker who therefore decides very early on to abandon the task, loses very little utility even in the absence of observability, and thus feels minimal guilt in leaving the requesting worker. The partner also possibly, might feel less empathy since they also were not going to get paid.

9.2.3 'Let's Play': The Role of Social Flow as an Incentive

When Wordsmith detects a worker's partner is about to exit the task, the worker can be prompted to send a message which says: 'Hi there, this is your partner, let's stay and tag a few more images'. This message is sent only if the requesting worker has been paid for tagging the requisite number of images, signifying a continuance for fun and pleasure. In the single worker mode of Wordsmith (Feyisetan et al., 2015b), we recorded scenarios where a worker who was paid to annotate eleven images, annotated over two hundred images. One of the factors that can be responsible for this, which is observed not only in games, but in everyday activities is *flow* or individual flow. The factors that make for individual flow, such as, clear goals, immediate feedback, and a balance between challenge and skills are present in the traditional single worker mode of Wordsmith, and are also present to a certain extent, in paid tasks. However, it might seem that the sense of personal control which characterizes individual flow is absent in the collaborative mode.

Social flow, on the other hand, could be identified in the collaborative theme of Wordsmith. Social flow builds on the experiences derived from solitary flow with additional conditions such as immediate and clear feedback from the task and group members, interdependence and cooperation, and conditions where the challenges are important to the whole group (Walker, 2010) and (Salanova et al., 2014). Social flow in the collaborative mode of Wordsmith could lead to shared absorption and engagement, and the desire to repeat the flow experience. The desire to repeat the experience represents the cost of exerting pressure on the partner who is about to exit the task. A worker therefore desiring to re-experience the flow might prompt their partner to stay for a few more rounds.

9.3 Experiment Design

In order to gain insight into the effect of collaboration in gamified paid microtasks, and specifically on the effects of social pressure, we performed a series of experiments on the Wordsmith platform, sourcing crowd workers from CrowdFlower. We carried out a within-subjects study in which a number of workers were recruited from a large pool, and required to annotate images either in the traditional single worker mode (**SP**) or the collaborative mode (**MP**). A crowd worker could participate in both studies. Within each task mode, workers were required to annotate a certain number of images in order to get paid: annotate 1 image (**LT** - *low threshold*) or annotate 11 images (**HT** - *high threshold*). Finally, in the collaborative, high-threshold mode (**MP-HT**), we carried out sub studies as follows: in one condition, a partner attempting to exit the task was allowed to, in another condition, a partner attempting to exit the task could be presented with a **please stay**, or **let's play** message.

9.3.1 Research Questions

We sought to answer the following research questions with our studies:

1. Does collaboration work as an effective model for paid microtask crowdsourcing?
2. Does the model work as quality assurance i.e., do answers converge faster?
3. What is the role of social incentives in collaborative crowdsourcing tasks?

9.3.2 Research Data

We utilised the image bank and keyword sets generated from the ESP game experiments (von Ahn and Dabbish, 2004). The dataset comprises of 100,000 images and about 1.4 million image tags. We screened out images with keywords which we deemed might be unsuitable for work environments, and selected a subset of images which had the highest number of keywords associated with them. We used the keywords as a sort of quasi ground truth to award bonus points within the task. In the **LT** (*low threshold*) conditions, we used a dataset size of 200 images, while in the **HT** (*high threshold*) conditions, we used a dataset of 2,200 images.

9.3.3 Worker Recruitment

We recruited all our participants primarily directly from CrowdFlower. We created separate tasks on CrowdFlower for the traditional and collaborative tasks. Crowd workers accessing the tasks were redirected to the Wordsmith platform where they carried out the annotation. In the traditional mode, we recruited 600 workers, assigning 3 workers to annotate each image, and in the collaborative mode, we also recruited 600 workers. For the **LT** (*low threshold*) conditions, workers were paid $0.02 each, while in the **HT** (*high threshold*) conditions, workers were paid $0.10 each.

9.3.4 Parameter 1: Wordsmith Task Mode

In our experiments, we studied the qualitative (match overlap with the ESP dataset keywords) and quantitative (number of labels generated) outputs between traditional single worker and collaborative multi-worker crowdsourcing modes.

Traditional (Single Worker) Mode In the traditional mode, workers individually tagged images without interaction or dependence on other workers. Workers could freely skip images which they were not interested in annotating.

Collaborative Mode In the collaborative mode, workers were paired up and required to guess keywords for the given image. When both participants matched on a set of labels, they were advanced to the next image. Even though the advancement of each worker was co-dependent on the effort of their partner, partners could be matched at different stages and therefore be eligible to receive their payoffs at different times.

9.3.5 Parameter 2: Task Threshold

We experimented with varying task thresholds which workers needed to attain before they got paid. The workers could then continue annotating further images at will after crossing the threshold.

Tag 1 Image Workers were required to tag one image to receive a payment of $0.02. This corresponded to advancing into the first level of the task.

Tag 11 Images Workers were required to tag eleven images to receive a payment of $0.10. This corresponded to advancing into the second level of the task.

9.3.6 Parameter 3: Social Incentives

We also experimented with two different forms in which a worker exerts social pressure on their partner in order to get them to remain in the task. When Wordsmith senses that a worker's partner is about to exit the task, the worker is alerted. They can then in turn opt to send a message requesting their partner to stay in the task. One of two messages could be sent:

Please Stay *(social pressure)*: If the requesting worker had tagged less than the requisite number of images to get paid, the following message gets sent to their partner: 'Please stay, I have not yet tagged x images, Don't leave me yet'.

Let's Play *(social flow)*: If the requesting worker had tagged more than the requisite number of images to get paid, the following message gets sent to their partner: 'Hi there, this is your partner, let's stay and tag a few more images'.

9.3.7 Experimental Conditions

Experiment 1 - Condition 1 (SP-LT): *Platform Mode: Traditional; Task: Tag 1 Image with at least 2 keywords; Source dataset size: 200 images; Workers: 600; Payment: $0.02.* In the first experiment, workers were required to tag 1 image with 2 keywords in the traditional mode.

Experiment 1 - Condition 2 (MP-LT): *Platform Mode: Collaborative; Task: Tag 1 Image with a paired partner with at least 2 keywords; Source dataset size: 200 images; Workers: 600; Payment: $0.02.* In the first experiment, paired workers were required to tag 1 image with 2 keywords in collaborative mode.

Experiment 2 - Condition 1 (SP-HT): *Platform Mode: Traditional; Task: Tag 11 Images with at least 2 keywords each; Source dataset size: 2,200 images; Workers: 600; Payment: $0.10.* In the experiment, workers were required to tag 11 images with 2 keywords each.

Experiment 2 - Condition 2 (MP-HT): *Platform Mode: Collaborative; Task: Tag 11 Images with a paired partner with at least 2 keywords each; Source dataset size: 2,200 images; Workers: 600; Payment: $0.10; Social incentives: None.* In the experiment, workers were required to tag 11 image with 2 keywords each in collaborative mode.

Experiment 2 - Condition 3 (MP-HT-SP): *Platform Mode: Collaborative; Task: Tag 11 Images with a paired partner with at least 2 keywords each; Source dataset size: 2,200 images; Workers: 600; Payment: $0.10; Social incentives: Please Stay and Let's Play.* In the experiment, workers were required to tag 11 image with 2 keywords each in collaborative mode and could be subject to social incentives.

Experiment Results					
	Low Threshold		High Threshold		
	Traditional	Collaborative	Traditional	Collaborative	Social Incentive
Total workers	402	365	514	499	508
Total tags	21,538	48,171	27,652	108,950	158,716
Unique images tagged	200	200	2,196	2,200	2,200
Inter-annotator agreement	29.44%	**34.55**%	14.26%	25.82%	**29.35**%
ESP tags agreement	41.26%	25.39%	43.96%	37.94%	40.11%
Avg images tagged / person	26.68 (SD=38.21)	9.77 (SD=13.23)	26.75 (SD=42.07)	25.05 (SD=17.92)	29.00 (SD=28.30)
Avg tags / person	53.57	**131.97**	53.80	218.34	**312.43**
Avg new tags / person	2.78 (1,117/402)	**8.69** (3,172/365)	1.80 (925/514)	11.83 (5,903/499)	**16.21** (8,236/508)

Table 9.1: *Experiment Results* - Summary of experiment results

9.4 Results

We recruited 600 workers for the single worker experiments modes and 600 workers for the collaborative experiments. In order to ascertain the veracity of our findings and the integrity of our results, we repeated the experiments twice and report here a contiguous set of results from one of the experiment runs. Table 9.1 summarises the results. The following sections describe tem along the lines of the different experimental condition parameters.

9.4.1 Wordsmith Task Mode: Traditional vs. Collaborative

We now present a summary of the results from the traditional and collaborative game modes from the experiments.

Participants
In the single worker mode, we put out a call for 600 workers. In the low threshold condition, we had a total of 416 annotating workers, while we had a total of 515 workers in the high threshold condition. Two factors contribute to the variance between these figures and the required value of 600, and between these figures compared with each other: (i) The single worker mode had a strong element of voluntary participation which allowed workers to skip images or ignore the task altogether, hence, fewer than the 600 participants recruited remained in the task and (ii) The high threshold condition paid more than the low threshold condition, which attracted more workers to the task. In the collaborative condition, we put out a call for 600 workers and received over $2,000$ workers assembling over $9,000$ teams and annotating over $4,500$ images across various experiment conditions.

Task Output
In the ***low threshold condition***, when workers were to annotate one image, we recorded a higher number of annotations (*avg images tagged / person*) in the traditional mode ($Avg. = 26.68$ images; $SD = 38.21$) than in the collaborative mode ($Avg. = 9.77$ images; $SD = 13.23$). The total number of tags (*total tags*) and average number of tags (*avg tags / person*) per person in the collaborative mode ($Total = 48,171$; $Avg. = 131.97$), was however significantly more than the total and average number of tags per worker in the traditional mode ($Total = 21,538$; $Avg. = 53.57$). The collaborative mode also generated more new tags, (*avg new tags / person*) than the traditional mode: new tags are labels which were not present in the ESP dataset but had a requisite threshold of inter-annotator agreement. There were $3,172$ new tags (8.69 new tags generated per worker) in the collaborative mode, versus $1,117$ new tags (2.78 per worker) in the traditional single worker mode.

Similarly, in the ***high threshold condition***, when workers were to annotate eleven images, we recorded a slightly higher number of annotations (*avg images tagged / person*) in the traditional mode (*Avg.* = 26.75 images; *SD* = 42.07) than in the collaborative mode (*Avg.* = 25.05 images; *SD* = 17.92 for the sub condition without social incentives). The total number of tags (*total tags*) and average number of tags (*avg tags / person*) per person in the single worker mode however (*Total* = 27,652; *Avg.* = 53.80), was orders of magnitude less than the total and average number of tags per worker in the collaborative mode (*Total* = 108,950; *Avg.* = 218.34). Even though workers annotated more individual images in the traditional as in the low threshold condition, they outputted on average, less tags than the collaborative mode. This is unsurprising, due to the fact that participants in the collaborative mode need to generate more guess tags to match with their partner. The collaborative mode also generated more new tags (*avg new tags / person*) than the traditional mode. There were 4,677 new tags (5.90 new tags generated per worker) in the collaborative mode, versus 1,166 new tags (2.26 per worker) in the traditional mode.

Comparing the task threshold conditions within the individual task modes reveals some further insights: within the traditional mode, there was no considerable difference in the average task output between the low (**SP-LT** - tag 1 image condition) at 26.68 images per person, and the high (**SP-HT** - tag 11 images conditions) at 26.75 images per person. Similarly, the output in the average number of tags within these conditions were almost the same with **SP-LT** at 53.57 tags per person, and **SP-HT** at 53.80 tags per person. This indicates that the task threshold does not essentially result in a significant change in the overall task output per worker, although it affects other dynamics e.g., the task covers a wider spectrum of image annotations in the high threshold. In the collaborative task however, we see a clear variance (with higher results in the high threshold) in the task output as a result of the task payment cutoff, with an average of 9.77 images and 131.97 tags per person in the low (**MP-LT**), and an average of 25.05 images and 218.34 tags per person in the high (**MP-HT**).

Task Quality

In the ***low threshold condition***, when workers were to annotate one image, they attained a higher inter-annotator agreement (Nowak and Rüger, 2010), in the collaborative mode (*A* = 34.55%), than in the traditional mode (*A* = 29.44%). We also computed the ESP agreement score, which was the agreement between a worker's annotation for a particular image, and the annotations for that image in the ESP dataset. The ESP agreement scores of each worker in the traditional (*ESP* = 41.26%) and collaborative modes (*ESP* = 25.39% when analysed individually), differed, with the traditional mode having a higher score. The collaborative mode however complemented this shortfall by having a higher number of new tags which were agreed on by workers from different teams, and which were not in the original ESP dataset.

Similarly, in the ***high threshold condition***, when workers were to annotate eleven images, they attained a higher inter-annotator agreement, in the collaborative mode (A = 25.82%), than in the traditional mode (A = 14.26%). The ESP agreement scores of each worker in the traditional (ESP = 43.96%) and collaborative modes (ESP = 37.94% when analysed individually), displayed less variance than in the low threshold condition. This is due to the fact that workers in the high threshold condition, collaboratively annotated on the average, as many images as participants in the low and high condition of the traditional setting (i.e., 25.05 images per person *vs* 26.68 and 26.75 images per person)

9.4.2 Wordsmith Incentives: With vs. Without Social Incentives

In this section, we present the results from the experiment conditions involving the multi-player game mode, with and without the social incentives of social pressure and social flow.

Task Output
In the collaborative, high threshold conditions (**MP-HT**), our results reveal that workers generated a higher number of total tags (*total tags*) and average tags per worker (*avg tags / person*) when they were subjected to social pressure (**MP-HT-SP**) from their partner (*Total* = $158,716$; *Avg.* = 312.43) than when they were not (*Total* = $108,950$; *Avg.* = 218.34). Workers annotated significantly more than the required number of images (29 images *vs* 11 required), exceeding the baseline averages (26 images) set in both the high and low threshold conditions of the traditional single worker setting. As a result of creating more tags overall, workers also generated a higher number of new labels between paired partners, and across teams annotating the same image (*avg new tags / person*) when their partners exerted social pressure and kept them in the task (*Total* = $8,236$ tags; *Avg.* = 16.21 tags) than when they didn't (*Total* = $5,903$ tags; *Avg.* = 11.83 tags).

Task Quality
Our results indicate a higher degree of agreement amongst the worker within the social pressure condition (**MP-HT-SP**) than those without (**MP-HT**). When compared with the ESP dataset, workers who could request their workers to stay and continue playing, achieved a slightly higher degree of quality, as a measure of agreement when each individual is assessed uniquely, (ESP = 40.11%) than those who couldn't (ESP = 37.94%). The inter-annotator scores between workers annotating the same image (with their partner and other workers who annotated the same image) were also higher with workers in the social pressure condition achieving a slightly greater inter-annotator score (A = 29.35% without social pressure; A = 25.05% with social pressure). This is also as a result of the social pressure condition incentivising a greater number of annotations.

Task Completion
When workers could request their partners to remain in the task, they were able to complete the task faster. As a result, the experiment condition that involved social pressure (**MP-HT-SP**) completed quicker (*Time* = 93hr 57min 31s) than the one (**MP-HT**) that didn't (*Time* = 131hr 57min 56s). Even though in both conditions, workers were explicitly alerted when their partner had left the task, workers who could not exert social pressure had to wait either for a new partner, or for their current partner to return after an extended break.

Team Switching
Workers in the social pressure condition formed fewer teams, stayed longer with their partners and switched teams less frequently when compared to those whose partners could leave at will. There were a total of $2,401$ teams created by 792 workers in the condition without social pressure (**MP-HT**), leading to an average switching rate of 3 teams per worker. In the social pressure condition (**MP-HT-SP**), workers formed fewer teams, $1,855$ by 810 workers, yielding resulting in a lower switching rate of 2.3 teams per worker.

9.4.3 Social Incentives: Please Stay vs. Let's Play

When Wordsmith senses that a worker's partner is about to exit the task, the worker is alerted. The worker can then request their partner to remain in the task. If the worker has tagged less than the requisite number of images in order to get paid, the worker can send a *please stay* request, else, the worker can send a *let's play* request. The receiving worker can then decide to stay (*i will stay*) or to leave the task (*i will go*).

Social Incentive Requests
From the results in Figure 9.4 with its accompanying table, we observe that workers are more likely to initiate a request of any kind when they have not been paid. When a worker has not yet been paid, they are more likely to request that their partner stay (*please stay request*) than permitting their partner to leave. After the workers had been paid, they were also more likely to request their partner to stay (*let's play request*) than permitting them to leave. The *please stay* requests (*Requests* = $1,023$) were used more frequently as a social incentive than the *let's play* request (*Requests* = 151), suggesting that workers are more inclined to put pressure on their partners when there is financial reward at stake than just fun. Figure 9.4 also reveals that some workers would actually release their partner to leave and wait to be connected to another partner. It indicates that on the average, as expected, fewer workers (20% *vs* 35%) who haven't been paid would opt for this option .

Figure 9.5 presents the distribution of *please stay* requests by workers at various image tag points. It illustrates clearly from the results, and from linear prediction trendlines,

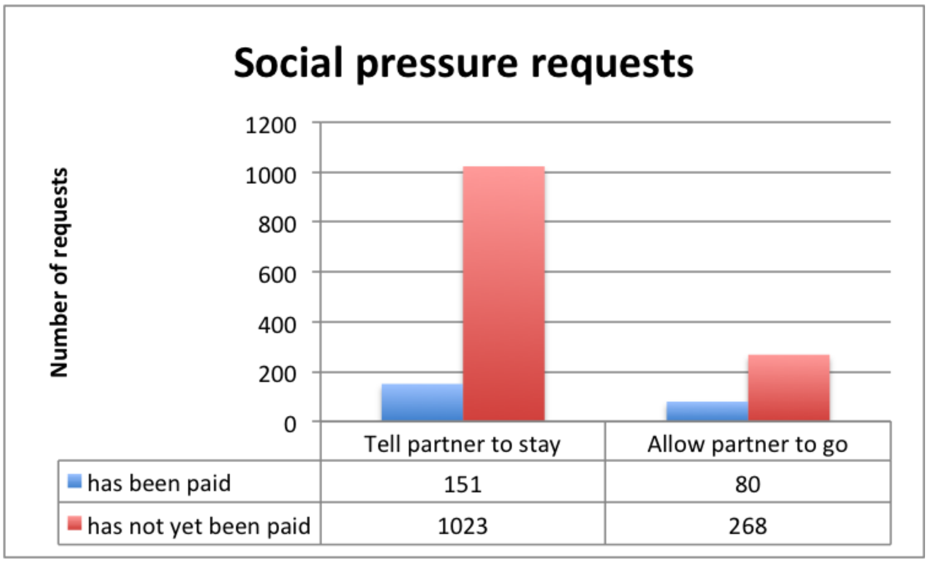


FIGURE 9.4: Social pressure requests made by workers before and after payment

that, the more a worker has invested into a task session, and as the reward of payment gets closer, a worker would be more reluctant to let their partner leave and would rather request that they stay. About 79% would request their partners to stay at the early stages of the task (the first 3 images), compared to 84% at the later stages (the last 3 images).

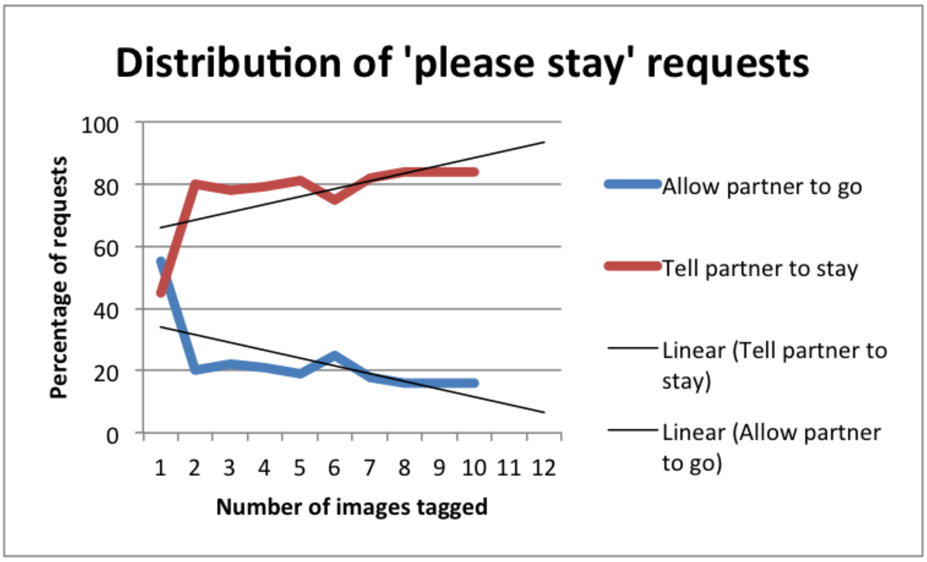
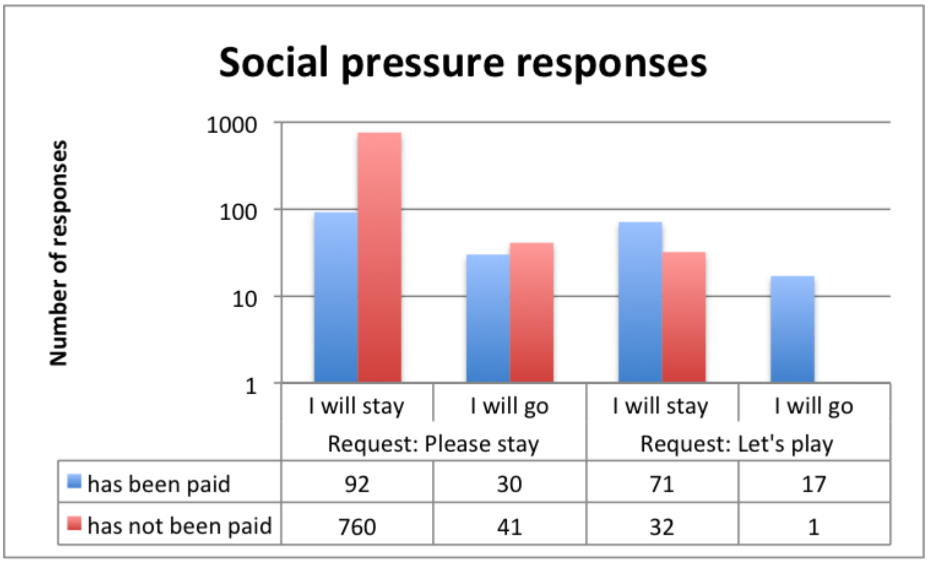


FIGURE 9.5: Distribution of *please stay* requests (with linear forecast trendlines) made by a worker after tagging specific number of images

Social Incentive Responses

Figure 9.6 summarises the results (in a logarithmic scale) of a worker's responses to both *please stay* and *let's play* requests. When a worker receives a *please stay* request (signifying that the requesting partner has not yet been paid), they can respond by choosing either to stay (*I will stay*) or to leave (*I will go*). The choice to stay or to leave also varies depending on whether the receiving worker has been paid or not. The results indicate that, a worker who has not been paid, receiving a *please stay* message from a fellow unpaid is more likely to stay, with 95% probability, than to exit the task (760 *vs* 41). This is in line with workers being incentivised by having shared circumstances (i.e., the need to both get paid), as stated by Kandel and Lazear (1992). Similarly, a worker receiving a *please stay* request from an unpaid worker, after they have been paid, is also likely to respond by staying, albeit, with a slightly less probability of 75% (92 *vs* 30). Furthermore, a worker receiving a *let's play* request (from a worker that has been paid) can also choose to stay or to leave, depending on whether the receiving worker has been paid or not. The results illustrate that, a worker who has not been paid, previously intending to exit the task, would almost certainly remain in the task after being sent a *let's play* message with 97% probability (32 *vs* 1). The result also reveals the response to social flow incentives: a worker who has been paid would return to continue playing with another worker with 80% certainty, even more likely than they would help a partner get paid (although, the results suggest that these requests occur less frequently). This is also another form of incentivisation by having shared circumstances (i.e., the desire to re-experience social flow).



	I will stay	I will go	I will stay	I will go
	Request: Please stay		Request: Let's play	
🟦 has been paid	92	30	71	17
🟥 has not been paid	760	41	32	1

Figure 9.6: Worker responses to *please stay* and *let's play* requests (on logarithmic scale)

Figure 9.6 also gives insights into when workers decide to leave their partners, despite receiving either a *please stay* or *let's play* request. The results reveal that, after receiving a *please stay* request from a worker who has not been paid, a receiving worker is more likely to leave if they have not been paid also. Hence they do not feel any guilt from leaving their partner hanging since they haven't been paid also. Similarly, after receiving a *let's play* request from a partner who has been paid, the receiving worker is more likely to decline the offer and choose to exit the task if they have also been paid.

Social Incentive Limits

Our results and Figure 9.7 indicate that paid microtask workers are more receptive to empathic social pressure incentives than social flow, although this is also as a result of this incentive being directed towards a larger base of actors. Workers responded positively more times to *please stay* requests ($Mean = 1.95$; $SD = 1.80$; $Max = 19$) than to *let's play* requests ($Mean = 1.16$; $SD = 0.52$; $Max = 4$).

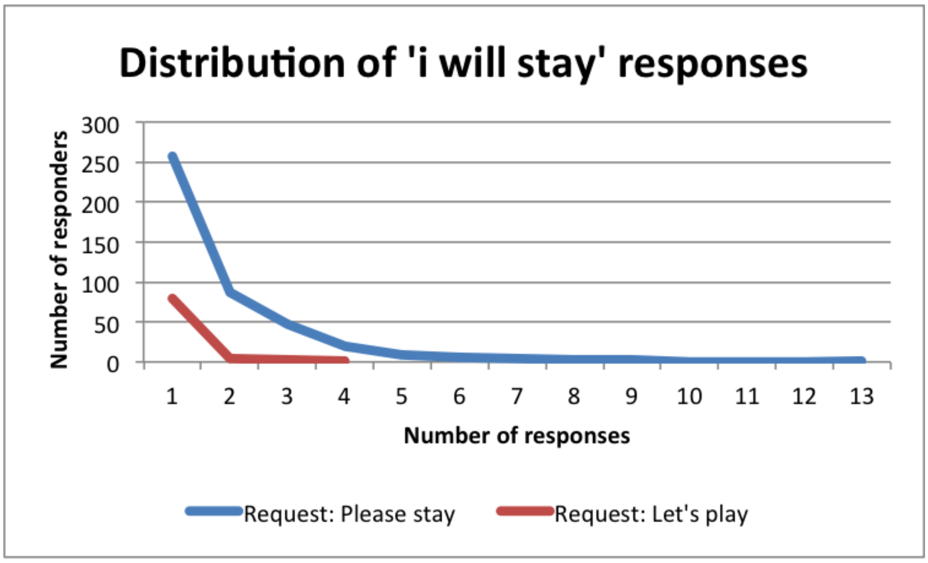


FIGURE 9.7: Distribution of *i will stay* responses: showing how many times a worker chose to stay after each request type

9.5 Discussion

Making Microtasks More Collaborative

Despite the individualistic nature of most paid microtask environments, our results reveal that collaborative multi-worker tasks are indeed attractive to members of this community. A worker posted on a discussion forum after one of our experiments: '*Hello everyone! lately I'm hooked on the multiplayer tasks, but are labor intensive*' (translated from Spanish). Our analysis indicate an improvement (represented by the total and average number tags as well as the number of new tags per person) in task output within the collaborative setting over the traditional (single worker) theme, at no additional costs. The results also suggest a higher quality output (from the inter-annotator agreement) by the fact that the collaborative mode requires consensus for task continuance. In all, workers were attracted to, willing to work collaboratively and exceeded baseline results in task output and quality. The findings suggest that adding support for more collaborative task design models into existing platforms, in addition to the community features which most of them have started to offer, may prove beneficial for requesters, and for workers. Such elements would also be beneficial for crowd training, allowing motivated participants to improve their performance. The experiments in Dow et al. (2012), for example, have looked at peer learning aspects, using workflows similar to ours; they suggest, for instance, to allow workers to be able to revise their answers based on the feedback they receive. While this might not be a practicable option in all scenarios, it could help train newcomers and the pair-based image labelling approach we introduced in this chapter could be one simple way to realize it. In Feyisetan et al. (2015b) the authors noted that one of the most effective furtherance incentives besides a financial reward was the ability to learn by studying the answers given by other people. While the feature they implemented did not offer much context (i.e., workers could see the answers of workers, but no explanations or validation), it suggests how popular these measures could be, and the effect reported was stronger for top contributors whom one can assume are driven by a desire to get better at the job.

Collaboration Must Pay Off

The benefits of collaborative participation in the image labelling task were more visible in the *high threshold* conditions. In the traditional mode, workers annotated, on average, the same number of images (and generated the same number of tags) in the high and low threshold conditions. In other words, without the restriction of partner agreements, the task threshold did not really make a difference. In the low task threshold condition of the collaborative mode, workers tagged more than the requisite number of images, nevertheless, this positive delta was not sufficient to match up to the individual freedom afforded in the traditional mode. The high task threshold on the other hand indicates in the collaborative setting, how the power of (and aspiration towards) social concordance, propped up by a higher payment cutoff can be leveraged to generate more and better results. Workers in this condition, initially motivated by the need to get paid, worked together to realise improved results. This finding contributes to the larger discussion around motivation and paid microtask crowdsourcing. Surveys such as Feyisetan et al. (2015b); Kaufmann et al. (2011); Mason and Watts (2010) have observed that financial incentives are just one, though important, part of a much more refined story of motivation of workers. The present work offers evidence on the effects of social pressure and social flow, but it also raises new questions regarding the implications of the findings for incentives design that take into account particular types of workers (e.g., top contributors vs. casual visitors).

Workers Behave Empathically

Social pressure incentives could be harnessed to attain speedy task completion and encourage empathic collaboration. Our analysis revealed that workers on realising that their partners have not reached the task threshold for payment, would be willing to annotate a few more images to help them get paid. This is in contrast to the individualistic thinking model which has been enshrined in traditional paid micro task platform settings. Our results demonstrated that paid workers would be willing, not only to work together, but to go the extra mile to ensure that their partner also gets paid. Workers respond not only to the need to help their partner get paid, they also respond to their partner's desire to continue annotating just for the fun of it. As noted earlier, while these results are encouraging, to develop a theory of incentives for paid microtask platforms, one would need additional experiments that take into account worker behaviour patterns, as well as other tasks and possibly more complex collaboration models.

9.6 Limitations

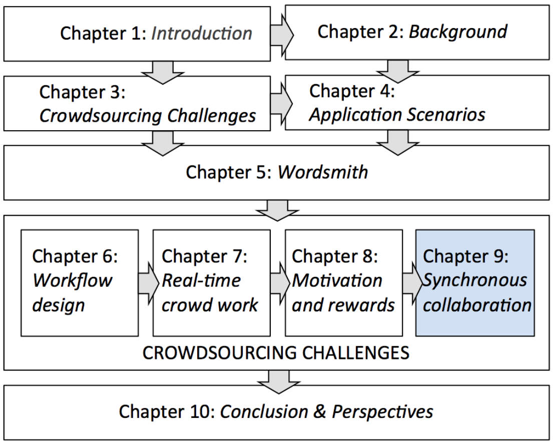
The experiments are run on image labelling tasks. While they could be easily adapted to similar task types, in particular to output-agreement ones (von Ahn and Dabbish, 2004), they are less suitable for settings in which a diversity of answers is sought. The same would apply for more complex tasks following, for instance, a 'fix-find-verify' (Bernstein et al., 2010) workflow. It would be interesting to explore how social pressure incentives would work in those settings, building on top of studies such as Kittur (2010).

The interplay between selected gamification features and the two types of incentives we have looked at has not yet been studied in detail. As such, the appeal the tasks had could be partially due to the use of gamification features, which are not common in paid microtask environments. Adding a single-worker condition to the experiments was meant to compensate for that, though a more thorough consideration of the alignment of specific game elements with social pressure and social flow has not yet been investigated.

9.7 Conclusion

The results shed light to our research questions by revealing collaboration as a viable means of undertaking paid microtask crowdsourcing. We demonstrated that paid microtask workers would be willing to undertake collaborative tasks, and generate significantly more useful output even at higher task requirements. Our results also show that this increased output does not result in a degradation, but rather an improvement of task quality. Furthermore we indicated that social incentives could be used to boost the performance of participants in this collaborative model. These results are in line with findings from GWAPs and multi-actor crowdsourcing systems, and could be used to inform the re-design of paid microtask platforms such as CrowdFlower and Mechanical Turk which do not integrate collaborative workflows as first-class citizens.

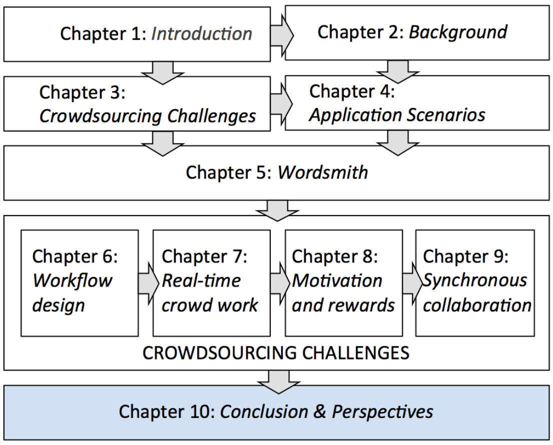
9.8 Summary


Chapter 1: Introduction
Chapter 2: Background
Chapter 3: Crowdsourcing Challenges
Chapter 4: Application Scenarios
Chapter 5: Wordsmith
Chapter 6: Workflow design
Chapter 7: Real-time crowd work
Chapter 8: Motivation and rewards
Chapter 9: Synchronous collaboration
CROWDSOURCING CHALLENGES
Chapter 10: Conclusion & Perspectives

This chapter concluded our survey of the four challenges in paid microtask crowdsourcing. It brought a culmination to our understanding of furtherance incentives as a broad medium of improving quality in paid microtasks. Our approach in this chapter expanded the single worker gamified mode presented previously in chapter 8 with an collaborative approach enhanced via social pressure and social flow as the furtherance incentive mechanisms.

Chapter 10

Conclusions and Perspectives


Chapter 1: *Introduction*
Chapter 2: *Background*
Chapter 3: *Crowdsourcing Challenges*
Chapter 4: *Application Scenarios*
Chapter 5: *Wordsmith*
Chapter 6: *Workflow design*
Chapter 7: *Real-time crowd work*
Chapter 8: *Motivation and rewards*
Chapter 9: *Synchronous collaboration*
CROWDSOURCING CHALLENGES
Chapter 10: *Conclusion & Perspectives*

In this chapter, we summarise the results and contributions of our work and its implications for crowdsourcing research. We highlight the findings from our studies into applying furtherance incentive techniques to address four crowdsourcing challenges, and present a list of future research areas along these lines. This brings to a close our work over the previous chapters and gives insight into pathways to create a viable and sustainable crowdsourcing ecosystem.

10.1 Summary

Crowdsourcing remains a mechanism with tremendous potential to grow economies by engaging a large workforce on demand and at scale. This translates into improved productivity and significant benefits for task requesters and crowd workers alike. Requesters gain task scalability, quick completion and lower price margins for their projects; while workers achieve additional income and a chance at social mobility. However, the obstacles to crowdsourcing, in particular, online paid microtask crowdsourcing, is equally well known and well studied. In addressing crowdsourcing as a socio-technical construct, we can also view its issues along these lenses: i.e., technical ones such as dealing with spam; and social issues requiring the evolution of crowdsourcing into a system which can support the future of (crowd) work.

This thesis focused on four core crowdsourcing challenges from a list of twelve highlighted by Kittur et al. (2013). The four crowdsourcing challenges discussed in this thesis were: (a) workflow design; (b) real-time crowd work; (c) motivation and rewards; and (d) synchronous collaboration. Although the thesis attempted to address the issues from a technical standpoint, two of the challenges had strong social components (motivation and rewards; and synchronous collaboration), while the other two had strong technical components (workflow design; and real-time crowd work). In order to adopt a well rounded methodology to tackle the challenges, over the chapters of this thesis we created a crowdsourcing tool, and progressively developed a methodology of addressing the selected issues. Our tool, *Wordsmith*, was a gamified microtask crowdsourcing platform which recruited crowd workers from existing marketplaces; while our methodology was the use of *furtherance incentives* which identified waning task participation and presented a means of re-engaging workers. We selected two broad application scenarios where crowdsourcing finds widespread applications: text processing and image labelling. These scenarios formed the basis of our experimentation in our core chapters.

Wordsmith served as a platform for carrying out microtasks, giving us the power to carry out the targeted experiments that gave us insights into our desired issues. Our work on *workflow design* presented an approach to improve microtask workflows by leveraging on task features and worker preferences. It also describes a way to potentially use our findings to afford for greater improvement by applying the task features to predict when workers would potentially quit a task, and then using the worker preferences as the furtherance incentive mechanism to re-engage workers.

Our work on *real-time crowd work* described an approach of using microtask contests to carry out judgements under tight constraints. This approach of using competitions naturally engenders worker exit given the varying utilities gained by individual workers during the contest. We presented a formal description into predicting when workers would quit a task (in this case, the contest) and described how we could then shore up appropriate incentives to workers who were about to quit the contest. This two pronged approach of detecting worker exit and presenting appropriate incentives form the framework for deploying furtherance incentives.

Our discussion on motivation and rewards completed our knowledge of furtherance incentives by putting the two building blocks together and presenting a theoretical model, and empirical evidence on the effectiveness of furtherance incentives. Our experiments on motivation and rewards highlighted some results crucial to the overall thesis: gamification, deployed appropriately, serves as a useful framework for improving worker engagement and task uptake; incentives deployed randomly presents a reasonable improvement to task performance than utilising plain task setups; and targeting furtherance incentives based on knowledge derived from worker behaviour, leads to significant engagement yielding higher quantity and quality metrics.

Finally, our work on synchronous collaboration built on our investigations into motivation and rewards by extending the Wordsmith platform to cater to paired collaborating workers. Our approach embraced a multiplayer methodology, which is prevalent in other crowdsourcing systems (such as GWAPs and citizen science projects) but has failed to attain a 'first class' role in crowdsourcing marketplaces. We also tested the full spectrum of furtherance incentives on this challenge area by using incentives designed around social constructs (social pressure and social flow).

Our results reinforced the role of furtherance incentives as a means of solving a wide range of crowdsourcing challenges which featured cash payments as the primary means of remuneration and incentivisation. The thesis also gave us first hand experience into crowdsourcing as a fully fledged socio-technical construct with challenges requiring as much a social touch as the technical. Crowd worker motivations especially, cover a wide spectrum which extends beyond the simplistic spectrum of 'fun and money', and, understanding how to keep them engaged is essential not just to yield superior returns for requesters, but also to create a more humane platform as crowdsourcing continues to take its place as the model for the future of work.

10.2 Future Work

From all our findings distilled into the past few chapters, we intend to continue our research path around the following directions:

Wordsmith: we intend to extend the research on our crowdsourcing platform primarily to accommodate other application scenarios. Currently, Wordsmith was designed to handle one of the six broad categories of crowdsourcing task types highlighted by Gadiraju et al. (2014) (i.e., content creation in the form of data annotation). Even within the narrow context of content creation, there remain other task types such as transcription, translation and text summarisation which could benefit from our furtherance incentive techniques. Additionally, we intend to carry out experiments into other broad task types including information finding, content access and analysis based tasks in order to gain an understanding into how crowd workers adapt or behave when faced with cognitively heavy tasks. Secondly, in order to make the tool more accessible for use by potentially non-technical users, we intend to integrate Wordsmith with configurable admin interfaces which would help with tasks such as to swap leaderboard strategies or select a different crowdsourcing marketplace (e.g., CrowdFlower vs Mechanical Turk).

Workflow design: in the light of our task and experiments into workflow design in microtasks, we intend to fill in several knowledge gaps to create a holistic framework that can be used to deploy end to end solutions for data annotation (as a starting point which could then be extended to other scenarios). First, we would devise automated approaches to determining when best to select human or machine capabilities – this is currently done in two separate strands of previous work which we would adequately integrate. Next, we would extend theory into practice by deploying workflow design tasks which use the entire furtherance incentives scheme: i.e., predicting when workers want to quit the task based on the implicit task and worker features, and then selecting furtherance incentives based on the type of tasks that the workers perform well. We would also carry out more research into *implicit named entities* which we uncovered as a medium to potentially target actual named entities in a sort of fashion carried out in anaphora detection research. Finally we would like to investigate the role of furtherance incentives in workflows involving experts as a final step to completing the machine – crowd – expert pipeline.

Real-time crowd work: in extending our research on real-time crowd work, we would like to carry out further study on the theoretical underpinnings of our work. Substantial theoretic research has been previously carried out on crowdsourcing contests in the field of economics and game theory and we would seek to extend such knowledge with our empirical experiments to design better systems. Specifically as a starting point, we would address one of the limitations of our current experiment setup: our task scenario of labelling entities relies on an assumption that we can correctly verify a worker's annotation in real-time – however, this is not foolproof and is a best-case event. However, some tasks, such as protein folding (such as in the Foldit project), can be easily carried out by the crowd and automatically verified by computers, making them a prime candidate for automatically verified real-time crowd work. Furthermore, we can model each worker state transition during the experiments using Markov chains to build a richer exit predictor. We would also experiment with varying payment methods as well as using other incentive mechanisms (such as gamification elements) in the contests.

Motivation and rewards: our future research into motivation and rewards serves as the central point where we would advance our knowledge on the core concepts of furtherance incentives. Among our ongoing efforts, we wish to better understand how and why the gamified incentives work in the ways and to the extent that they do. Our current study of furtherance incentives target workers at the potential point of impending exit. In particular, we believe that furtherance incentives could be more effective if carefully distributed within the game mechanics so that they appear at appropriate intervals when motivation begins to wane, not only after the participant has initiated an attempt to leave. Second, we wish to improve the probabilistic model to take into account other macro realities that exist in crowdsourcing e.g., other aspects of players performance, task history and demographic, and to understand the ways in estimating appropriate rewards. In particular, we wish to run further experiments to determine whether incentives are more effective for particular group settings than for others, or for workers at particular skill levels or task completion histories.

Synchronous collaboration: we would like to extend our work on synchronous collaboration to feature ideas present in other groupsourcing platforms. Some of these include affording for larger teams (currently limited to paired teams of two) and creating a platform for team dynamicity featuring self-selection and automatic group assignment. Research into dynamic multi-actor teams would give us further insight into ideal team sizes for different task types which would feature full scale collaboration and interaction between team members. We would also experiment with different payment structures as the team size grows – this is to observe for, and prevent social loafing which can creep in as the number of workers increase. Furthermore, we would carry out cross studies between synchronous collaboration and real-time crowd work, and how the former can be used to facilitate the later. Finally, we would investigate how crowd IQ, speed of decision-making and task consensus changes with crowd size and other macro and micro features.

This thesis focused on four out of an initial selection of twelve crowdsourcing challenges. As we seek to apply our methods to other broad crowdsourcing task types, we would also investigate the suitability for some of the other challenges.

References

S. Ahmad, A. Battle, Z. Malkani, and S. Kamvar. The jabberwocky programming environment for structured social computing. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, pages 53–64. ACM, 2011.

H. Ali-Hassan and D. Nevo. Identifying social computing dimensions: A multidimensional scaling study. In *Proceedings of the 30th International Conference on Information Systems*, page 148. Association for Information Systems, 2009.

M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, (2):76–81, 2013.

A. Anagnostopoulos, L. Becchetti, C. Castillo, A. Gionis, and S. Leonardi. Online team formation in social networks. In *Proceedings of the 21st International Conference on World Wide Web*, pages 839–848. ACM, 2012.

N. Archak. Money, glory and cheap talk: analyzing strategic behavior of contestants in simultaneous crowdsourcing contests on topcoder. com. In *Proceedings of the 19th International Conference on World Wide Web*, pages 21–30. ACM, 2010.

N. Archak and A. Sundararajan. Optimal design of crowdsourcing contests. In *Proceedings of the 30th International Conference on Information Systems*, page 200. Association for Information Systems, 2009.

L. Aroyo and C. Welty. Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *Proceedings of the 5th Annual ACM Web Science Conference*, 2013.

L. Aroyo and C. Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, mar 2015.

G. Attardi et al. Phratris – A Phrase Annotation Game. *Demo presented at Incentives for Semantics Game Idea Challenge*, 2010.

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *DBpedia: A Nucleus for a Web of Open Data.* Springer, 2007.

A. E. C. Basave, A. Varga, M. Rowe, M. Stankovic, and A. Dadzie. Making Sense of Microposts (# MSM2013) Concept Extraction Challenge. In *#MSM2013*, pages 1–15, 2013.

J. Bennett and S. Lanning. The Netflix Prize. In *Proceedings of KDD Cup and Workshop*, volume 2007, page 35, 2007.

T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth. Who's in the picture. *Advances in Neural Information Processing Systems*, 17:137–144, 2005.

T. Berners-Lee and M. Fischetti. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor.* Harper San Francisco, 1st edition, 1999.

M. S. Bernstein, J. Brandt, R. C. Miller, and D. R. Karger. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, pages 33–42. ACM, 2011.

M. S. Bernstein, D. R. Karger, R. C. Miller, and J. Brandt. Analytic methods for optimizing realtime crowdsourcing. *arXiv preprint arXiv:1204.2995*, 2012.

M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich. Soylent: a word processor with a crowd inside. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology*, pages 313–322. ACM, 2010.

P. K. Bhowmick, P. Mitra, and A. Basu. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 58–65. Association for Computational Linguistics, 2008.

J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, et al. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology*, pages 333–342. ACM, 2010.

J. Bitzer, W. Schrettl, and P. J. H. Schröder. Intrinsic motivation in open source software development. *Journal of Comparative Economics*, 35(1):160–169, 2007.

I. Blohm and J. M. Leimeister. Gamification: Design of IT-Based Enhancing Services for Motivational Support and Behavioral Change. *Business & Information Systems Engineering*, pages 275–278, 2013.

I. Bogost. Persuasive games: exploitationware. *Gamasutra, May*, 3, 2011.

I. Bogost. Why gamification is bullshit 2. *The gameful world: Approaches, issues, applications*, 65, 2015.

K. Bontcheva, L. Derczynski, and I. Roberts. Crowdsourcing named entity recognition and entity linking corpora. *Handbook of Linguistic Annotation*, 2014a.

K. Bontcheva, I. Roberts, L. Derczynski, and D. Rout. The gate crowdsourcing plugin: Crowdsourcing annotated corpora made easy. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 97–100, 2014b.

K. J. Boudreau, N. Lacetera, and K. R. Lakhani. Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Science*, 57(5):843–863, 2011.

A. Bozzon, M. Brambilla, S. Ceri, M. Silvestri, and G. Vesci. Choosing the right crowd: expert finding in social networks. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 637–648, 2013.

D. C. Brabham. *Crowdsourcing*. MIT Press, 2013.

J. Bragg, A. Kolobov, M. Mausam, and D. S. Weld. Parallel task routing for crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2014.

K. Braunschweig, M. Thiele, J. Eberius, and W. Lehner. Enhancing named entity extraction by effectively incorporating the crowd. In *BTW Workshops*, pages 181–195, 2013.

M. J. Brzozowski, T. Sandholm, and T. Hogg. Effects of feedback and peer pressure on contributions to enterprise social media. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, pages 61–70. ACM, 2009.

M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1): 3–5, 2011.

R. Burkett. An Alternative Framework for Agent Recruitment: From MICE to RASCLS. In *Studies in Intelligence*, volume 57 of *1*, pages 7–17, mar 2013.

A. Calvó-Armengol and M. O. Jackson. Peer pressure. *Journal of the European Economic Association*, 8(1):62–89, 2010.

C. F. Camerer and R. M. Hogarth. The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19 (1-3):7–42, 1999.

A. Carvalho, S. Dimitrov, and K. Larson. The output-agreement method induces honest behavior in the presence of social projection. *Newsletter of the ACM Special Interest Group on E-commerce*, 13(1):77–81, 2014.

R. Cavallo and S. Jain. Efficient crowdsourcing contests. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, pages 677–686. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

J. Chamberlain. Groupsourcing: Distributed problem solving using social networks. In *Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing*, 2014a.

J. Chamberlain. The Annotation-validation (AV) Model: Rewarding Contribution Using Retrospective Agreement. In *Proceedings of the 1st International Workshop on Gamification for Information Retrieval*, pages 12–16. ACM, 2014b.

J. Chamberlain, M. Poesio, and U. Kruschwitz. Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics' 08)*, pages 42–49, 2008.

R. L. II Charles. Social network email filtering, March 2 2010. US Patent 7,673,003.

S. Chawla, J. D. Hartline, and B. Sivan. Optimal crowdsourcing contests. *Games and Economic Behavior*, 2015.

Y. Cheng, Z. Chen, J. Wang, A. Agrawal, and A. Choudhary. Bootstrapping active name disambiguation with crowdsourcing. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*, pages 1213–1216. ACM, 2013.

H. Chesbrough. *Open innovation: The new imperative for creating and profiting from technology.* Harvard Business School Press, 2006.

N. A. Chinchor. Overview of MUC-7/MET-2. 1998.

J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

T. Cohn and L. Specia. Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 32–42. ACL, 2013.

S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, et al. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, 2010.

M. Csikszentmihalyi. *Flow: The psychology of optimal experience*, volume 41. Harper-Perennial New York, 1991.

K. Daido. Risk-averse agents with peer pressure. *Applied Economics Letters*, 11(6): 383–386, 2004.

K. Daido. Peer pressure and incentives. *Bulletin of Economic Research*, 58(1):51–60, 2006.

A. Das, U. Burman, B. Ar, and S. Bandyopadhyay. NER from Tweets: SRI-JU System. In *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts'*, page 62, 2013.

R. Dawson and S. Bynghall. *Getting Results from Crowds.* Advanced Human Technologies, 2012.

V. De Boer, M. Hildebrand, L. Aroyo, P. De Leenheer, C. Dijkshoorn, B. Tesfa, and G. Schreiber. Nichesourcing: harnessing the power of crowds of experts. In *Knowledge Engineering and Knowledge Management*, pages 16–20. Springer, 2012.

M. J. A. N. de Caritat et al. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* L'imprimerie royale, 1785.

A. G. S. de Herrera, A. Foncubierta-Rodríguez, D. Markonis, R. Schaer, and H. Müller. Crowdsourcing for medical image classification. *Swiss Medical Informatics*, 30, 2014.

D. M. de Oliveira, A. H. F. Laender, A. Veloso, and A. S. da Silva. FS-NER: A Lightweight Filter-stream Approach to Named Entity Recognition on Twitter Data. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 597–604, 2013.

J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, 51(1):107–113, 2008.

E. Dechenaux, D. Kovenock, and R. M. Sheremeta. A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics*, pages 1–61, 2014.

E. L. Deci, R. Koestner, and R. M. Ryan. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6):627, 1999.

E. L. Deci and R. M. Ryan. *Intrinsic motivation.* Wiley Online Library, 1975.

E. L. Deci and R. M. Ryan. Cognitive Evaluation Theory. In *Intrinsic Motivation and Self-Determination in Human Behavior*, pages 43–85. Springer, 1985a.

E. L. Deci and R. M. Ryan. *Intrinsic Motivation and Self-Determination in Human Behavior.* Perspectives in Social Psychology. Springer, 1985b.

G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st International Conference on World Wide Web*, pages 469–478. ACM, 2012.

G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Large-scale linked data integration using probabilistic reasoning and crowdsourcing. *The VLDB Journal*, 22(5):665–687, October 2013.

J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49, 2015.

S. Deterding. Gamification: Designing for Motivation. *interactions*, 19(4):14–17, 2012.

S. Deterding, D. Dixon, R. Khaled, and L. Nacke. From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pages 9–15. ACM, 2011a.

S. Deterding, R. Khaled, L. Nacke, and D. Dixon. Gamification: Toward a definition. In *CHI 2011 Gamification Workshop Proceedings*, pages 12–15, 2011b.

D. E. Difallah, M. Catasta, G. Demartini, P. G. Ipeirotis, and P. Cudré-Mauroux. The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. In *Proceedings of the 24th International Conference on World Wide Web*, pages 238–247, 2015.

D. E. Difallah, G. Demartini, and P. Cudré-Mauroux. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. *CrowdSearch 2012 Workshop at WWW*, pages 26–30, 2012.

D. E. Difallah, G. Demartini, and P. Cudré-Mauroux. Pick-a-crowd: Tell me what you like, and i'll tell you what to do - a crowdsourcing platform for personalized human intelligence task assignment based on social networks. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 367–377, may 2013.

D. DiPalantino and M. Vojnovic. Crowdsourcing and all-pay auctions. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, pages 119–128. ACM, 2009.

A. Doan, R. Ramakrishnan, and A. Y. Halevy. Crowdsourcing Systems on the World-Wide Web. *Communications of the ACM*, 54(4):86–96, April 2011.

G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M. Weischedel. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In *Proceedings of the Language Resources and Evaluation Conference*, 2004.

S. Dow, A. Kulkarni, S. Klemmer, and B. Hartmann. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 1013–1022. ACM, 2012.

A. Dumitrache, L. Aroyo, C. Welty, R. Sips, and A. Levas. Dr. Detective: combining gamication techniques and crowdsourcing to create a gold standard in medical text. In *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web*, pages 16–31. CEUR-WS. org, 2013.

C. Eickhoff, C. G. Harris, A. P. de Vries, and P. Srinivasan. Quality through flow and immersion: Gamifying crowdsourced relevance assessments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 871–880. ACM, 2012.

N. Eisenberg and P. A. Miller. The relation of empathy to prosocial and related behaviors. *Psychological Bulletin*, 101(1):91, 1987.

R. Eisenberger, W. D. Pierce, and J. Cameron. Effects of reward on intrinsic motivationNegative, neutral, and positive: Comment on Deci, Koestner, and Ryan (1999). *Psychological Bulletin*, 1999.

L. Erickson, C. Baughman, and P. Distribution. *Top Secret Rosies: The Female Computers of World War II.* PBS Distribution, 2010.

L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

A. Felstiner. Working the crowd: employment and labor law in the crowdsourcing industry. *Berkeley Journal of Employment and Labor Law*, pages 143–203, 2011.

O. Feyisetan, M. Luczak-Rösch, E. Simperl, R. Tinati, and N. Shadbolt. Towards Hybrid NER: A Study of Content and Crowdsourcing-Related Performance Factors. In *The Semantic Web. Latest Advances and New Domains*, pages 525–540. Springer, 2015a.

O. Feyisetan, E. Simperl, R. Tinati, M. Luczak-Rösch, and N. Shadbolt. Quick-and-clean extraction of linked data entities from microblogs. In *Proceedings of the 10th International Conference on Semantic Systems*, pages 5–12. ACM, 2014.

O. Feyisetan, E. Simperl, M. Van Kleek, and N. Shadbolt. Improving paid microtasks through gamification and adaptive furtherance incentives. In *Proceedings of the 24th International Conference on World Wide Web*, pages 333–343, 2015b.

Oluwaseyi Feyisetan and Elena Simperl. Please stay vs lets play: Social pressure incentives in paid collaborative crowdsourcing. In *International Conference on Web Engineering*, pages 405–412. Springer, 2016.

T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88. Association for Computational Linguistics, 2010.

B. Forrest. Technology saves lives in haiti. *Forbes.com*, 2010.

D. A. Forsyth and J. Ponce. A modern approach. *Computer Vision: A Modern Approach*, pages 88–101, 2003.

B. Frei. Paid crowdsourcing. *Current State & Progress toward Mainstream Business Use, Smartsheet. com Report, Smartsheet. com*, 9, 2009.

B. S. Frey and R. Jegen. Motivation crowding theory. *Journal of Economic Surveys*, 15 (5):589–611, 2001.

H. Fromreide, D. Hovy, and A. Søgaard. *Crowdsourcing and annotating NER for Twitter #drift.* European Language Resources Distribution Agency, 2014.

D. Fudenberg and J. Tirole. A theory of exit in duopoly. *Econometrica: Journal of the Econometric Society*, pages 943–960, 1986.

U. Gadiraju, R. Kawase, and S. Dietze. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, pages 218–223. ACM, 2014.

F. Galton. The most suitable proportion between the value of 1st and second prizes. *Biometrika*, pages 385–399, 1902.

F. Galton. Vox Populi (The Wisdom of Crowds). *Nature*, 75:450–51, 1907.

H. Gao, G. Barbier, and R. Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, (3):10–14, 2011a.

H. Gao, X. Wang, G. Barbier, and H. Liu. Promoting coordination for disaster relief – from crowdsourcing to coordination. In *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 197–204. Springer, 2011b.

D. Geiger, S. Seedorf, T. Schulze, R. C. Nickerson, and M. Schader. Managing the crowd: Towards a taxonomy of crowdsourcing processes. In *Proceedings of the 17th Americas Conference on Information Systems*, 2011.

Y. Genc, W. A. Mason, and J. V. Nickerson. Classifying Short Messages using Collaborative Knowledge Bases: Reading Wikipedia to Understand Twitter. In *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts'*, pages 50–53, 2013.

M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, 40(16):6266–6282, 2013.

A. Ghosh and P. Hummel. Cardinal contests. In *Proceedings of the 24th International Conference on World Wide Web*, pages 377–387, 2015.

A. Ghosh and P. McAfee. Crowdsourcing with endogenous entry. In *Proceedings of the 21st International Conference on World Wide Web*, pages 999–1008. ACM, 2012.

A. Go, L. Huang, and R. Bhayani. Twitter sentiment analysis. *Entropy*, 17, 2009.

M. F. Goodchild and J. A. Glennon. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3):231–241, 2010.

N. Green, P. Breimyer, V. Kumar, and N. F. Samatova. PackPlay: Mining semantic data in collaborative games. In *Proceedings of the 4th Linguistic Annotation Workshop*, pages 227–234. Association for Computational Linguistics, 2010.

S. Greenberg and R. Bohnet. Group sketch: A multi-user sketchpad for geographically-distributed small groups. In *Proceedings of Graphics Interface*. University of Calgary, 1991.

G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007.

J. Hamari, J. Koivisto, and H. Sarsa. Does gamification work? – a literature review of empirical studies on gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences*, pages 3025–3034. IEEE, 2014.

B. Han and T. Baldwin. Lexical Normalisation of Short Text Messages: Makn Sens a #Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 368–378. ACL, 2011.

C. Harris. You're hired! an examination of crowdsourcing incentive models in human resource tasks. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM)*, WSDM 2011, pages 15–18, 2011.

C. G.. Harris and P. Srinivasan. Comparing Crowd-Based, Game-Based, and Machine-Based Approaches in Initial Query and Query Refinement Tasks. In *Advances in Information Retrieval*, volume 7814, pages 495–506. Springer Berlin Heidelberg, 2013.

A. Hars and S. Ou. Working for free? motivations of participating in open source projects. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, pages 9–pp. IEEE, 2001.

J. Heinzelman and C. Waters. *Crowdsourcing crisis information in disaster-affected Haiti*. US Institute of Peace, 2010.

G. Hertel, S. Niedner, and S. Herrmann. Motivation of software developers in open source projects: an internet-based survey of contributors to the linux kernel. *Research Policy*, 32(7):1159–1177, 2003.

A. Hinze, R. Heese, M. Luczak-Rösch, and A. Paschke. Semantic enrichment by non-experts: usability of manual annotation tools. In *Proceedings of the International Semantic Web Conference*, pages 165–181. Springer, 2012a.

A. Hinze, R. Heese, A. Schlegel, and M. Luczak-Rösch. User-defined semantic enrichment of full-text documents: Experiences and lessons learned. In *Theory and Practice of Digital Libraries*, pages 209–214. Springer, 2012b.

B. Hladká, J. Mírovskỳ, and J. Kohout. An attractive game with the document:(im) possible? *The Prague Bulletin of Mathematical Linguistics*, 96:5–26, 2011.

C. Ho, A. Slivkins, S. Suri, and J. W. Vaughan. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, pages 419–429, 2015.

J. J. Horton and L. B. Chilton. The Labor Economics of Paid Crowdsourcing. In *Proceedings of the 11th ACM Conference on Electronic Commerce*, pages 209–218. ACM, 2010.

E. Hovy and J. Lavid. Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. *International Journal of Translation*, 22(1):13–36, 2010.

J. Howe. The rise of crowdsourcing. *Wired Magazine*, 14(6):1–4, 2006.

J. Howe. *Crowdsourcing: How the power of the crowd is driving the future of business.* Random House, 2008.

M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K. Ma. Breaking news on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2751–2754. ACM, 2012.

G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

S. Huang and W. Fu. Don't hide in the crowd!: increasing social transparency between peer workers improves crowdsourcing outcomes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 621–630. ACM, 2013.

P. G. Ipeirotis. Analyzing the Amazon Mechanical Turk Marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16–21, 2010a.

P. G. Ipeirotis. Demographics of Mechanical Turk. In *NYU Working Paper No. CEDER-10-01*, 2010b.

P. G. Ipeirotis. Mechanical turk vs odesk: My experiences. *Retrieved from http://www.behind-the-enemy-lines.com/2012/02/mturk-vs-odesk-my-experiences.html (accessed November 1, 2014)*, 2012.

L. C. Irani and M. Silberman. Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 611–620. ACM, 2013.

S. A. Jackson and M. Csikszentmihalyi. *Flow in sports*. Human Kinetics, 1999.

A. H. Jadidinejad. Unsupervised Information Extraction using BabelNet and DBpedia. In *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts'*, pages 54–56, 2013.

D. Jurgens and R. Navigli. It's All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation. *Transactions of the Association for Computational Linguistics (TACL)*, 2:449–464, 2014.

E. Kandel and E. P. Lazear. Peer pressure and partnerships. *Journal of political Economy*, pages 801–817, 1992.

S. J. Karau and K. D. Williams. Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65(4):681, 1993.

N. Kaufmann, T. Schulze, and D. Veit. More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk. In *Proceedings of the 17th Americas Conference on Information Systems*, pages 1–11, 2011.

S. Kerr. On the folly of rewarding a, while hoping for b. *Academy of Management Journal*, 18(4):769–783, 1975.

J. S. Kim, M. J. Greene, A. Zlateski, K. Lee, M. Richardson, S. C. Turaga, M. Purcaro, M. Balkam, A. Robinson, B. F. Behabadi, et al. Space-time wiring specificity supports direction selectivity in the retina. *Nature*, 509(7500):331, 2014.

A. Kittur. Crowdsourcing, collaboration and creativity. *ACM Crossroads*, 17(2):22–26, 2010.

A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 453–456, 2008.

A. Kittur, S. Khamkar, P. André, and R. Kraut. Crowdweaver: visually managing complex crowd work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 1033–1036. ACM, 2012.

A. Kittur, B. Lee, and R. E. Kraut. Coordination in collective intelligence: the role of team structure and task interdependence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1495–1504. ACM, 2009.

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 ACM Conference on Computer Supported Cooperative Work*, pages 1301–1318. ACM, 2013.

A. Kittur, B. Smus, S. Khamkar, and R. E. Kraut. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, pages 43–52. ACM, 2011.

A. Kobren, C. H. Tan, P. G. Ipeirotis, and E. Gabrilovich. Getting More for Less: Optimized Crowdsourcing with Dynamic Tasks and Goals. In *Proceedings of the 24th International Conference on World Wide Web*, pages 592–602, 2015.

E. Kouloumpis, T. Wilson, and J. Moore. Twitter sentiment analysis: The good the bad and the omg. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. Citeseer, 2011.

R. Kraut, M. L. Maher, J. Olson, T. W. Malone, P. Pirolli, and J. C. Thomas. Scientific Foundations: A Case for Technology-Mediated Social-Participation Theory. *Computer*, 43(11):22–28, 2010.

R. E. Kraut, P. Resnick, S. Kiesler, M. Burke, Y. Chen, N. Kittur, J. Konstan, Y. Ren, and J. Riedl. *Building successful online communities: Evidence-based social design*. MIT Press, 2012.

V. Krishna and J. Morgan. An Analysis of the War of Attrition and the All-Pay Auction. *Journal of Economic Theory*, 72(2):343–362, 1997.

P. Kucherbaev, F. Daniel, S. Tranquillini, and M. Marchese. Crowdsourcing processes: a survey of approaches and opportunities. *IEEE Internet Computing*, 20(2):50–56, 2016.

A. Kulkarni, M. Can, and B. Hartmann. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 1003–1012. ACM, 2012.

S. Kumar, G. Barbier, M. A. Abbasi, and H. Liu. TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

S. Kuznetsov. Motivations of contributors to wikipedia. *ACM SIGCAS Computers and Society*, 36(2):1, 2006.

H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World wide web*, pages 591–600. ACM, 2010.

K. Lakhani and R. G. Wolf. Why hackers do what they do: Understanding motivation and effort in free/open source software projects. *Perspectives on Free and Open Source Software*, 2005.

D. Laniado and P. Mika. Making Sense of Twitter. In *Proceedings of the 9th International Semantic Web Conference*, pages 470–485. Springer, 2010.

W. Lasecki, C. Miller, A. Sadilek, A. Abumoussa, D. Borrello, R. Kushalnagar, and J. Bigham. Real-time captioning by groups of non-experts. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, pages 23–34. ACM, 2012a.

W. S. Lasecki, C. Homan, and J. P. Bigham. Architecting real-time crowd-powered systems. *Human Computation*, 1(1), 2014.

W. S. Lasecki, C. D. Miller, and J. P. Bigham. Warping time for more effective real-time crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2033–2036. ACM, 2013a.

W. S. Lasecki, K. I. Murray, S. White, R. C. Miller, and J. P. Bigham. Real-time crowd control of existing interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, pages 23–32. ACM, 2011.

W. S. Lasecki, Y. C. Song, H. Kautz, and J. P. Bigham. Real-time crowd labeling for deployable activity recognition. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 1203–1212. ACM, 2013b.

W. S. Lasecki, S. C. White, K. I. Murray, and J. P. Bigham. Crowd memory: Learning in the collective. In *Proceedings of Collective Intelligence*, 2012b.

B. Latane, K. Williams, and S. Harkins. Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, 37 (6):822, 1979.

E. Law and L. von Ahn. Input-agreement: A New Mechanism for Collecting Data Using Human Computation Games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1197–1206. ACM, 2009.

N. Lawson, K. Eustice, M. Perkowitz, and M. Yetisgen-Yildiz. Annotating large email datasets for named entity recognition with Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 71–79. Association for Computational Linguistics, 2010.

E. P. Lazear and S. Rosen. Rank-order tournaments as optimum labor contracts. *The Journal of Political Economy*, 89(5):841–864, 1981.

M. R. Lepper, D. Greene, and R. E. Nisbett. Undermining children's intrinsic interest with extrinsic reward: A test of the 'overjustification' hypothesis. *Journal of Personality and Social Psychology*, 28(1):129, 1973.

G. M. Levitt. *The Turk, Chess Automation.* McFarland & Company, Incorporated Publishers, 2000.

G. Little, L. B. Chilton, M. Goldman, and R. C. Miller. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology*, pages 57–66. ACM, 2010.

T. X. Liu, J. Yang, L. A. Adamic, and Y. Chen. Crowdsourcing with all-pay auctions: A field experiment on taskcn. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–4, 2011a.

X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 359–367. Association for Computational Linguistics, 2011b.

A. Loparev, W. S. Lasecki, K. I. Murray, and J. P. Bigham. Introducing shared character control to existing video games. 2014.

M. Luczak-Rösch and R. Heese. Linked data authoring for non-experts. In *Proceedings of the WWW 2009 Workshop on Linked Data on the Web*, 2009.

M. A. Luengo-Oroz, A. Arranz, and J. Frean. Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears. *Journal of Medical Internet Research*, 14(6):e167, 2012.

T. W. Malone, R. Laubacher, and C. Dellarocas. The Collective Intelligence Genome. *MIT Sloan Management Review*, 51(3):21–31, 2010.

A. Mao, E. Kamar, Y. Chen, E. Horvitz, M. E. Schwamb, C. J. Lintott, and A. M. Smith. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In B. Hartman and E. Horvitz, editors, *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing.* AAAI, 2013a.

A. Mao, E. Kamar, and E. Horvitz. Why stop now? predicting worker engagement in online crowdsourcing. In *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing*, 2013b.

M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

M. Marrero, S. Sanchez-Cuadrado, J. M. Lara, and G. Andreadakis. Evaluation of named entity extraction systems. *Advances in Computational Linguistics, Research in Computing Science*, 41:47–58, 2009.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the 8th IEEE International Conference on Computer Vision*, volume 2, pages 416–423. IEEE, 2001.

D. Martin, B. V. Hanrahan, J. O'Neill, and N. Gupta. Being a turker. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 224–235. ACM, 2014.

W. Mason and D. J. Watts. Financial incentives and the performance of crowds. *ACM SIGKDD Explorations Newsletter*, 11(2):100–108, 2010.

M. J. Mataric. Designing emergent behaviors: From local interactions to collective intelligence. In *Proceedings of the 2nd International Conference on From Animals to Animats 2 : Simulation of Adaptive Behavior: Simulation of Adaptive Behavior*, pages 432–441. MIT Press, 1993.

S. Mavandadi, S. Dimitrov, S. Feng, F. Yu, U. Sikora, O. Yaglidere, S. Padmanabhan, K. Nielsen, and A. Ozcan. Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study. *PloS One*, 7(5):37245, 2012.

D. Maynard, K. Bontcheva, and H. Cunningham. Towards a semantic extraction of named entities. *Recent Advances in Natural Language Processing*, 2003.

J. McGonigal. *Reality Is Broken: Why Games Make Us Better and How They Can Change the World.* The Penguin Group, 2011.

C. Mellström and M. Johannesson. Crowding out in blood donation: was titmuss right? *Journal of the European Economic Association*, 6(4):845–863, 2008.

P. N. Mendes, D. Weissenborn, and C. Hokamp. DBpedia Spotlight at the MSM2013 Challenge. In *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts'*, pages 57–61, 2013.

P. Michelucci. *Handbook of Human Computation.* Springer, 2013.

E. Minkov, R. C. Wang, and W. W. Cohen. Extracting personal names from email: applying named entity recognition to informal text. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 443–450. Association for Computational Linguistics, 2005.

C. A. Mockros and M. Csikszentmihalyi. The social construction of creative lives. In *The Systems Model of Creativity*, pages 127–160. Springer, 2014.

A. Mohnen, K. Pokorny, and D. Sliwka. Transparency, inequity aversion, and the dynamics of peer pressure in teams: Theory and evidence. *Journal of Labor Economics*, 26(4):693–720, 2008.

B. Moldovanu and A. Sela. The optimal allocation of prizes in contests. *American Economic Review*, pages 542–558, 2001.

B. Moldovanu, A. Sela, and X. Shi. Carrots and sticks: prizes and punishments in contests. *Economic Inquiry*, 50(2):453–462, 2012.

R. Morris. Crowdsourcing workshop: the emergence of affective crowdsourcing. In *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems*, 2011.

D. Morrison, S. Marchand-Maillet, and É. Bruno. Tagcaptcha: annotating images with captchas. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 44–45. ACM, 2009.

O. Muñoz-García, A. García-Silva, and O. Corcho. Towards Concept Identification using a Knowledge-Intensive Approach. In *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts'*, pages 45–49, 2013.

E. Musk. All our patent are belong to you. blog, June 2014. URL: http://www.teslamotors.com/blog/all-our-patent-are-belong-you. Accessed: 2015-03-13.

D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. 2013.

M. J. Nelson. Soviet and american precursors to the gamification of work. In *Proceeding of the 16th International Academic MindTrek Conference*, pages 23–26. ACM, 2012.

J. Nickerson. Crowd work and collective learning. *Technology-Enhanced Professional Learning: Routledge, Forthcoming*, 2013.

B. Norrander. The attrition game: Initial resources, initial contests and the exit of candidates during the us presidential primary season. *British Journal of Political Science*, 36(03):487–507, 2006.

O. Nov. What motivates wikipedians? *Communications of the ACM*, 50(11):60–64, 2007.

S. Nowak and S. Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, pages 557–566. ACM, 2010.

J. Ortmann, M. Limbu, D. Wang, and T. Kauppinen. Crowdsourcing linked open data for disaster management. In *Proceedings of the Terra Cognita Workshop on Foundations, Technologies and Applications of the Geospatial Web in conjunction with the ISWC*, pages 11–22. Citeseer, 2011.

A. Pak and P. Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *In Proceedings of the 7th International Conference on Language Resources and Evaluation*, volume 10, pages 1320–1326, 2010.

M. Parameswaran and A. B. Whinston. Research issues in social computing. *Journal of the Association for Information Systems*, 8(6):336, 2007.

E. Pavlick, M. Post, A. Irvine, D. Kachaev, and C. Callison-Burch. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92, 2014.

S. Petrovic, M. Osborne, R. McCreadie, C. Macdonald, I. Ounis, and L. Shrimpton. Can twitter replace newswire for breaking news? In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 2013.

S. Phuvipadawat and T. Murata. Breaking news detection and tracking in twitter. In *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 120–123. IEEE, 2010.

B. Plank, D. Hovy, and A. Søgaard. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of EACL*, 2014.

M. Poesio, J. Chamberlain, and U. Kruschwitz. Phrase detectives. *Ide, N., and Pustejovsky, J., eds*, 2015.

A. J. Quinn and B. B. Bederson. Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1403–1412. ACM, 2011.

J. Raddick, C. Lintott, S. Bamford, K. Land, D. Locksmith, P. Murray, B. Nichol, K. Schawinski, A. Slosar, A. Szalay, D. Thomas, J. Vandenberg, and D. Andreescu. Galaxy Zoo: Motivations of Citizen Scientists. In *Bulletin of the American Astronomical Society*, volume 40, page 240, May 2008.

M. J. Raddick, G. Bracey, K. Carney, G. Gyuk, K. Borne, J. Wallin, S. Jacoby, and A. Planetarium. Citizen science: status and research directions for the coming decade. *AGB Stars and Related Phenomenastro 2010: The Astronomy and Astrophysics Decadal Survey*, page 46P, 2009.

S. Rafaeli and Y. Ariel. Online motivational factors: Incentives for participation and contribution in wikipedia.(2008). 2008.

W. Rafelsberger and A. Scharl. Games with a purpose for social networking platforms. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, pages 193–198. ACM, 2009.

R. Ramanath, M. Choudhury, K. Bali, and R. S. Roy. Crowd prefers the middle path: A new iaa metric for crowdsourcing reveals turker biases in query segmentation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1713–1722, 2013.

P. Resnick and H. R. Varian. Recommender systems. *Communications of the ACM*, 40 (3):56–58, 1997.

A. Ritter, S. Clark, O. Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

G. Rizzo and R. Troncy. NERD: Evaluating Named Entity Recognition Tools in the Web of Data. pages 1–16, 2011.

J. A. Roberts, I. Hann, and S. A. Slaughter. Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the apache projects. *Management Science*, 52(7):984–999, 2006.

D. Robertson and F. Giunchiglia. Programming the social computer. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987), 2013.

P. Rogers. The cognitive psychology of lottery gambling: A theoretical review. *Journal of Gambling Studies*, 14(2):111–134, 1998.

J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In L. A. Adamic, R. A. Baeza-Yates, and S. Counts, editors, *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. The AAAI Press, 2011.

M. Rokicki, S. Chelaru, S. Zerr, and S. Siersdorfer. Competitive game designs for improving the cost effectiveness of crowdsourcing. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1469–1478. ACM, 2014.

M. Rokicki, S. Zerr, and S. Siersdorfer. Groupsourcing: Team competition designs for crowdsourcing. In *Proceedings of the 24th International Conference on World Wide Web*, pages 906–915, 2015.

J. Ross, L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2863–2872. ACM, 2010.

B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2008.

R. M. Ryan and E. L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68, 2000.

M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, 2014.

M. Sabou, A. Scharl, and F. Michael. Crowdsourced knowledge acquisition: Towards hybrid-genre workflows. *International Journal on Semantic Web and Information Systems*, 9(3):14–41, 2013.

S. Sachidanandan, P. Sambaturu, and K. Karlapalem. NERTUW: Named Entity Recognition on Tweets using Wikipedia. In *Proceedings of the Concept Extraction Challenge at the Workshop on 'Making Sense of Microposts'*, pages 67–70, 2013.

H. Saif, Y. He, and H. Alani. Semantic Sentiment Analysis of Twitter. In *Proceedings of the 11th International Conference on The Semantic Web*, pages 508–524. Springer, 2012.

M. Salanova, A. M. Rodríguez-Sánchez, W. B. Schaufeli, and E. Cifre. Flowing together: A longitudinal study of collective efficacy and collective flow among workgroups. *The Journal of Psychology*, 148(4):435–455, 2014.

J. Schroer and G. Hertel. Voluntary engagement in an open web-based encyclopedia: Wikipedians and why they do it. *Media Psychology*, 12(1):96–120, 2009.

K. Seaborn and D. I. Fels. Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, 74:14–31, 2015.

E. Seltzer and D. Mahmoudi. Citizen participation, open innovation, and crowdsourcing: Challenges and opportunities for planning. *Journal of Planning Literature*, pages 1–16, 2012.

N. R. Shadbolt, D. A. Smith, E. Simperl, M. Van Kleek, Y. Yang, and W. Hall. Towards a classification framework for social machines. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 905–912, 2013.

A. D. Shaw, J. J. Horton, and D. L. Chen. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, pages 275–284. ACM, 2011.

V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622. ACM, 2008.

E. Simperl, R. Cuel, and M. Stein. Incentive-centric semantic web application engineering. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 3(1):1–117, 2013.

V. K. Singh, R. Piryani, A. Uddin, and P. Waila. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *Proceedings of the International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, pages 712–717. IEEE, 2013.

P. R. Smart and N. R. Shadbolt. Social machines. In M. Khosrow-Pour, editor, *Encyclopedia of Information Science and Technology*. IGI Global, January 2014.

P. R. Smart, E. Simperl, and N. Shadbolt. *A Taxonomic Framework for Social Machines*. Springer, 2014.

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

A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. 2008.

K. Starbird. Digital volunteerism during disaster: Crowdsourcing information processing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 7–12, 2011.

E. A. Stohr and J. L. Zhao. Workflow Automation: Overview and Research Issues. *Information Systems Frontiers*, 3(3):281–296, 2001.

J. Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, May 2004.

J. C. Tang, M. Cebrian, N. A. Giacobe, H. Kim, T. Kim, and D. B. Wickert. Reflecting on the DARPA red balloon challenge. *Communications of the ACM*, 54(4):78–85, 2011.

J. P. Tangney and R. L. Dearing. *Shame and guilt*. Guilford Press, 2003.

S. Thaler, E. Simperl, and S. Wölger. An experiment in comparing human-computation techniques. *IEEE Internet Computing*, 16(5):52–58, 2012.

S. Thaler, K. Siorpaes, D. Mear, E. Simperl, and C. Goodman. Seafish: a game for collaborative and visual image annotation and interlinking. In *The Semantic Web: Research and Applications*, pages 466–470. Springer, 2011.

R. Tinati, M. Van Kleek, E. Simperl, M. Luczak-Rösch, R. Simpson, and N. Shadbolt. Designing for citizen data analysis: A cross-sectional case study of a multi-domain citizen science platform. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4069–4078. ACM, 2015.

K. Tjong, F. Erik, and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL*, pages 142–147. Association for Computational Linguistics, 2003.

B. Trushkowsky, T. Kraska, M. J. Franklin, and P. Sarkar. Crowdsourced enumeration queries. In *29th International Conference on Data Engineering (ICDE)*, pages 673–684. IEEE, 2013.

R. Usbeck, A. Ngomo, M. Röder, D. Gerber, S. A. Coelho, S. Auer, and A. Both. Agdistis-graph-based disambiguation of named entities using linked data. In *Proceedings of the 13th International Semantic Web Conference*, pages 457–471. Springer, 2014.

W. M. van Der Aalst, A. H. Ter Hofstede, B. Kiepuszewski, and A. P. Barros. Workflow patterns. *Distributed and Parallel Databases*, 14(1):5–51, 2003.

F. Vis. Twitter as a reporting tool for breaking news: Journalists tweeting the 2011 UK riots. *Digital Journalism*, 1(1):27–47, 2013.

L. Von Ahn. *Human Computation*. PhD thesis, 2005. AAI3205378.

L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326. ACM, 2004.

L. von Ahn and L. Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, August 2008.

L. Von Ahn, S. Ginosar, M. Kedia, and M. Blum. Improving image search with phetch. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages IV–1209. IEEE, 2007.

L. Von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 55–64. ACM, 2006.

R. Voyer, V. Nygaard, W. Fitzgerald, and H. Copperman. A hybrid model for annotating named entity training corpora. In *Proceedings of the 4th Linguistic Annotation Workshop*, pages 243–246. Association for Computational Linguistics, 2010.

C. Wah. Crowdsourcing and its applications in computer vision. *University of California, San Diego*, 2006.

C. J. Walker. Experiencing flow: Is doing it together better than doing it alone? *The Journal of Positive Psychology*, 5(1):3–11, 2010.

A. Wang, C. D. V. Hoang, and M. Kan. Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, 47(1):9–31, 2013.

A. Weiss. The Power of Collective Intelligence. *netWorker*, 9(3):16–23, September 2005.

R. W. White. Motivation reconsidered: the concept of competence. *Psychological Review*, 66(5):297, 1959.

H. Xie and J. C. S. Lui. Modeling crowdsourcing systems: Design and analysis of incentive mechanism and rating system. *ACM SIGMETRICS Performance Evaluation Review*, 42(2):52–54, September 2014.

J. Yan and S. Yu. Magic bullet: a dual-purpose computer game. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 32–33. ACM, 2009.

D. Yang, G. Xue, X. Fang, and J. Tang. Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing. In *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, pages 173–184. ACM, 2012.

J. Yang, L. A. Adamic, and M. S. Ackerman. Crowdsourcing and knowledge sharing: strategic user behavior on taskcn. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 246–255. ACM, 2008.

Y. Ye and K. Kishida. Toward an understanding of the motivation of open source software developers. In *Proceedings of the 25th International Conference on Software Engineering*, pages 419–429. IEEE, 2003.

M. Yetisgen-Yildiz, I. Solti, F. Xia, and S. R. Halgrim. Preliminary experience with Amazon's Mechanical Turk for annotating medical named entities. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 180–183. Association for Computational Linguistics, 2010.

M. Yin, Y. Chen, and Y. Sun. The effects of performance-contingent financial incentives in online labor markets. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA.*, 2013.

L. Yu and J. V. Nickerson. Cooks or cobblers?: crowd creativity through combination. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1393–1402. ACM, 2011.

M. Yuen, I. King, and K. Leung. A survey of crowdsourcing systems. In *Proceedings of the Third International Conference on Privacy, Security, Risk and Trust (PASSAT)*, pages 766–773. IEEE, 2011.

H. Zheng, D. Li, and W. Hou. Task design, motivation, and participation in crowdsourcing contests. *International Journal of Electronic Commerce*, 15(4):57–88, 2011.

G. Zichermann. Gamification has issues, but they aren't the ones everyone focuses on, 2011.

G. Zichermann and C. Cunningham. *Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps.* 2011.

M. Zook, M. Graham, T. Shelton, and S. Gorman. Volunteered geographic information and crowdsourcing disaster relief: a case study of the haitian earthquake. *World Medical & Health Policy*, 2(2):7–33, 2010.

A. Zubiaga, M. Liakata, R. Procter, K. Bontcheva, and P. Tolmie. Crowdsourcing the annotation of rumourous conversations in social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 347–353, 2015.