Fall 2013

Applying human computation methods to information science

Christopher Glenn Harris
University of Iowa

APPLYING HUMAN COMPUTATION METHODS TO INFORMATION SCIENCE

by

Christopher Glenn Harris

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Informatics
in the Graduate College of
The University of Iowa

December 2013

Thesis Supervisor: Professor Padmini Srinivasan

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

PH.D. THESIS

_____

This is to certify that the Ph.D. thesis of

Christopher Glenn Harris

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Informatics at the December 2013 graduation.

Thesis Committee: _____
Padmini Srinivasan, Thesis Supervisor

_____
Alberto Segre

_____
Walter Vispoel

_____
Juan Pablo Hourcade

_____
Gautam Pant

To my wife, Shuhong and my children, Brady, Ronan and Miliani

ACKNOWLEDGMENTS

I wish to thank my thesis supervisor Padmini Srinivasan for her guidance and vision as I developed and executed this research the emerging field. In particular, I appreciate her direction, vision and ideas on how human computation can contribute to a new paradigm for accomplishing tasks in many different disciplines, including Information Science. I also wish to thank the others on my thesis committee for their direction on how to proceed in my research during the most challenging times, their prudent guidance, and their careful feedback on my ideas in human computation. I also wish to thank those students and faculty in the University of Iowa Text Retrieval group, many of whom have contributed in different ways to support this research, including providing useful feedback on new ideas, sharing new perspectives on long-standing problems in Information Retrieval, assisting me with research design, and reviewing research papers. Last but certainly not least, I wish to thank my wife and children who have patiently encouraged me along the way.

ABSTRACT

Human Computation methods such as crowdsourcing and games with a purpose (GWAP) have each recently drawn considerable attention for their ability to synergize the strengths of people and technology to accomplish tasks that are challenging for either to do well alone. Despite this increased attention, much of this transformation has been focused on a few selected areas of information science. This thesis contributes to the field of human computation as it applies to areas of information science, particularly information retrieval (IR). We begin by discussing the merits and limitations of applying crowdsourcing and game-based approaches to information science. We then develop a framework that examines the value of using crowdsourcing and game mechanisms to each step of an IR model. We identify three areas of the IR model that our framework indicates are likely to benefit from the application of human computation methods: acronym identification and resolution, relevance assessment, and query formulation. We conduct experiments that employ human computation methods, evaluate the benefits of these methods and report our findings. We conclude that employing human computation methods such as crowdsourcing and games, can improve the accuracy of many tasks currently being done by machine methods alone. We demonstrate that the best results can be achieved when human computation methods augment computer-based IR processes, providing an extra level of skills, abilities, and knowledge that computers cannot easily replicate.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

INTRODUCTION

Crowdsourcing, as a human computation approach, has been defined as the act of taking a job traditionally performed by a designated agent (such as an employee or a student) and making it available to an undefined, generally large group of people in the form of an open call (Howe, 2006). Although crowdsourcing is not a new concept, the recent rise in attention placed on crowdsourcing is due to several factors, including the ubiquity of the Internet, the need to increase performance on tasks that computers cannot do well (such as relevance judgments, geo-tagging, and image annotations), the improved worldwide reach of micropayment methods, the cost of hiring experts, as well as the disparity of global economic labor demand and tight local labor restrictions. In fact, the large worker supply, little regulation, and low labor costs provide crowdsourcing's strongest advantages (O. Alonso & Lease, 2011a).

As with many new technologies, the early days of crowdsourcing have primarily focused on the areas with the greatest need: repetitive, single-purpose tasks designed around a single objective, such as image classification, video annotation, form-based data entry, optical character recognition, translation, and document proofreading. However, as crowdsourcing begins to mature, it has begun a transformation, creating fascinating new opportunities for leveraging real-time human computation for a range of diverse and complex tasks, such as providing quality assurance in the peer review process in biological research (Meyer *et. al.*, 2012), providing a detailed check of submitted expense receipts by the U.K.'s Members of Parliament (Davies, 2009), or analyzing spectrograms and sounds of whales in an attempt to decipher them (Thompson, 2012).

Crowdsourcing is also changing academic research methodologies, allowing for greater responsiveness, task effectiveness, and affordability. This is seen most clearly in

areas such as engineering, computer science, linguistics, and psychology where crowdsourcing is enabling new forms of data collection and user studies.

Crowdsourcing has expanded academic research in new directions as well. Since Amazon.com introduced Mechanical Turk[1] in 2005, this mechanism has become a phenomenon in academic research reaching across many disciplines. Consider language translation: an examination of the number of articles returned by a Google Scholar search for "language translation" shows slow but steady growth over the period 2006-2012. A similar search combining crowdsourcing-related terms with the search phrase "language translation" shows exponential growth over the same period (see Figure 1 for data collected on June 25, 2013). Although during the period 2006-2012 research in language translation appeared to expand slowly, an increasing share of this research involved crowdsourcing.

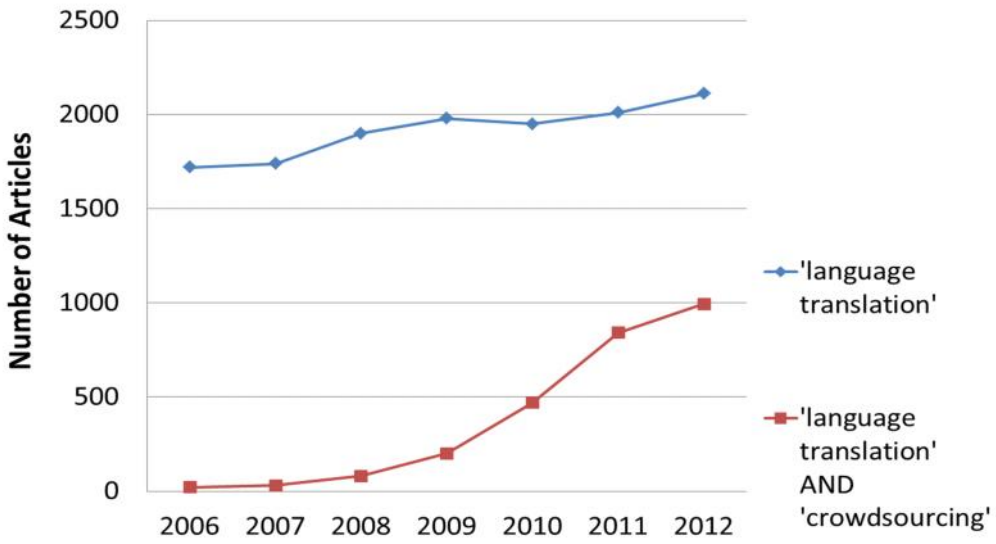

Figure 1: Number of Google Scholar articles containing specific terms by year, 2006-2012

[1] http://www.mturk.com

In effect crowdsourcing has not only facilitated other research, but it has become a research area of its own. There are new challenges and new opportunities in this intersection of people and technology. Design components such as human factors and human-computer interaction are essential to successful crowdsourcing tasks, as are consideration of psychology, law, ethics, systems testing, etc. – all of which need to be integrated with existing principles of software architecture and application design. This may in turn provide functionality or accuracy not previously thought achievable.

Online games offer a well-established form of entertainment; demographic studies show that the number of hours that people participate in games per week is increasing. Game play is no longer limited to a specific player demographic, and games have been gaining widespread acceptance as a tool to accomplish a specific goal, such as a supplement to learning. This last point – the use of Games With A Purpose (GWAP) – is also our focus in this research. It is becoming easier to cajole players to accomplish a task while being entertained, which has been behind much of the recent development and use of games. GWAP also have the ability to make mundane tasks engaging, provide a lower cost-per-task than even crowdsourcing can claim, and, which compared with crowdsourcing, has shown the potential of reducing noisy inputs. Although computers have made phenomenal advances over the past half century, they still do not possess the conceptual intelligence or perceptual capabilities most humans take for granted. If human brainpower is treated as part of a large distributed network, each node may be able to perform a tiny fragment of a large computation. The trouble is, unlike algorithms, humans need an inducement or incentive to become part of a collective computation. Thus, the allure of online games serves as a seductive method for encouraging people to participate in this process.

Since 2005, when MTurk was first established, a significant proportion of the tasks involving crowdsourcing and/or GWAP have related to IR. Tasks include relevance judgments, classification, labeling and annotating text and images. Since document relevance is a highly-subjective task, taking a majority vote across multiple assessors is often

done. Crowdsourcing and GWAP address this well as judgments can be obtained inexpensively, quickly and – when the relevance task is designed well – with high quality. Likewise, classification tasks also rely on the abilities of the worker to correctly group items, so the diversity obtained through crowdsourcing or GWAP is an advantage. Finally, labeling and annotating text is often a tedious exercise, making it prone to errors. Crowdsourcing and GWAP can address these potential errors by permitting additional labels to be generated through several members of the crowd, and spreading out the workload to more people, reducing the tedium. An example of this is LabelMe, a tool that provides image labels; several workers annotate a single image, and these crowdworker-supplied annotations are then merged.

The use of a huge pool of participants allows large-scale tedious tasks to become scalable. In their 2004 paper, von Ahn and Dabbish – the creators of the ESP Game – hypothesize that 5,000 individuals playing their game continuously could label each of the 425 million images indexed by Google within 31 days (Luis von Ahn & Dabbish, 2004). Indeed, popular online game websites, such as Yahoo! Games, feature many games that often have more than 5,000 concurrent players. The 31-day estimate is for labeling each of the images with a single keyword. In six months, the authors indicate that each image could be labeled with a minimum of six keywords. Although the number of images on Google has increased significantly since 2004, the use of a game-based format involving human participants is still considered the best way to accomplish this large-scale image-labeling task.

We observe that crowdsourcing and GWAP are exciting frontiers of computing that involve a mix of different disciplines. Our research focuses on how crowdsourcing and GWAP can improve tasks related to the Information Retrieval subarea of Information Science; however, there are many other disciplines where potential benefits of applying these mechanisms are also possible, such as was attained in collaborative filtering algorithms with the 2009 Netflix prize for improving the accuracy of movie recommendations.

Although research on the utility of crowdsourcing and game mechanisms has grown rapidly in recent years, few attempts have been made to classify the types of tasks these mechanisms address. For the few that have been created, these classifications do not attempt to apply themselves to real-world scenarios, such as to IR systems. Moreover, although IR models have been discussed extensively in the literature none of these discussions directly address how crowd and game-based mechanisms can add value to the Steps required for the IR processes found in our model.

Likewise, in the literature to date, there have been no empirical studies that examine the use of crowdsourcing and serious games under similar conditions. Those articles that mention these human computation approaches rely on literature review to identify some differences between these two human computation genres, but they do not provide the context in which they differ or indicate the degree in which one approach affects the results as compared with the other. Our research compares modes of interaction (game interfaces, non-game interfaces) and participant type (crowdworkers, students) to allow us to the difference in the effect it has on established performance measures.

In this thesis, we wish to examine the following three questions. First, can we use human computation to improve performance in IS tasks, particularly those involving IR? Second, which human computation factors are most appropriate for a given IS task? Third, in which situations does the augmentation of an algorithmic approach with human computation make sense?

1.1 An Information Retrieval System

Our goal is to study crowdsourcing and GWAP in the context of Information Science (IS), specifically in Information Retrieval (IR). Hence we start with a model of IR, shown in Figure 2. Similar IR models have been developed by others (e.g., Lancaster and Warner have developed a model analogous to ours (Lancaster & Warner, 1993). Our model illustrates a typical process for building an IR application and using it. Although there are other aspects to

IR not shown, (e.g., document translation); it forms a reasonable starting point for examining the fit of crowdsourcing and GWAP to IR tasks. The model depicts the process of establishing and running an IR system as a sequence of typical Steps. Steps 1-6 (on the left-hand side) designate the IR system design and implementation (preparatory stage) and Steps 7-12 (on the right hand side) designate the user query processing (interactive stage).

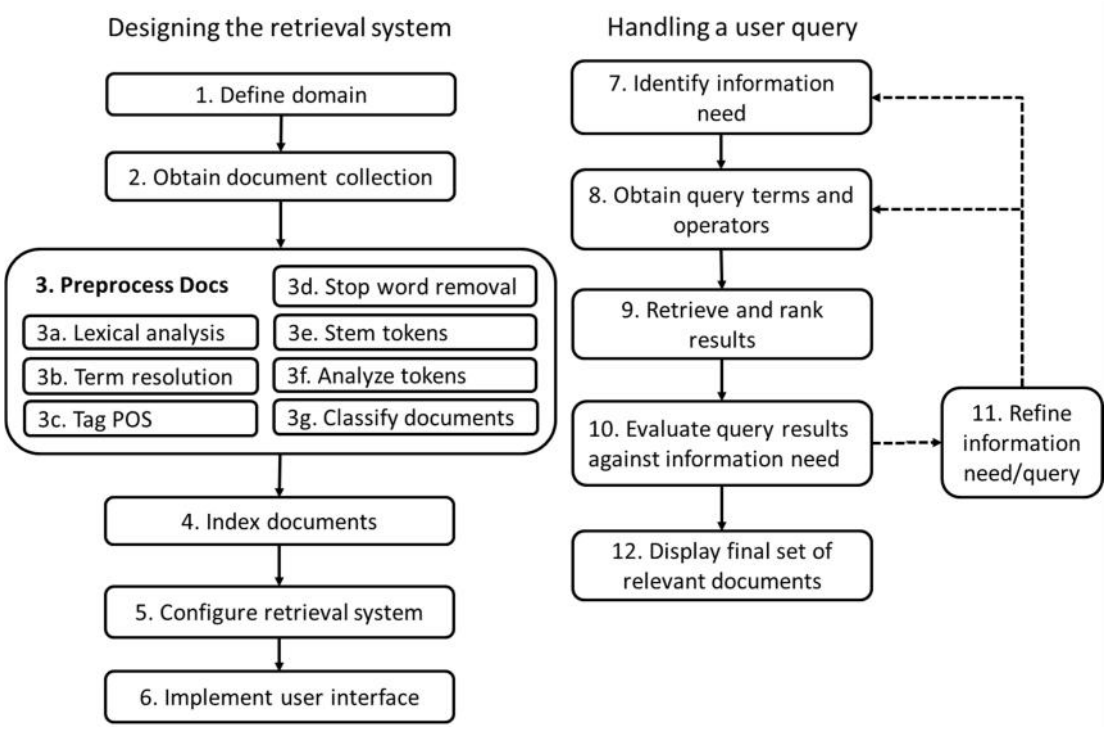

Figure 2: A process oriented information retrieval model

We begin with the left-hand side of Figure 2. First, the domain, which contains the boundaries of the collection, is defined. Next, we identify and obtain an appropriate document collection. Then in Step 3, we preprocess the collection's documents prior to indexing. This typically includes lexical analysis, term resolution (resolving abbreviations and acronyms), part-of-speech tagging, stop word removal, stemming and analysis of tokens, and document classification.

Table 1: Step description, task objective and expected output from each step of our IR model.

Step No	Step Description	Task Objective	Expected Output
1	Define domain	Define the scope of information for retrieval tasks	A definition of the domain (boundaries and nature of the contents) applicable to task
2	Obtain document collection	Determine the documents to be indexed and available for retrieval. Web spidering, web link analysis, etc., could be done at this stage.	List of documents or document sources, and, if applicable, a set of inter-document links.
3	Preprocess documents	Provide data enrichment	Contains multiple objectives and expected outputs. See Table 2 for additional details.
4	Index documents	Create an index for all documents	A set of indexed documents
5	Configure retrieval system	Determine the best search strategies and parameters for an anticipated set of tasks	Best search strategy and parameters within a chosen retrieval system for the collection
6	Implement user interface	Have an interface that users can use to meet their needs	A usable interface to all anticipated user group interaction with retrieval documents
7	Identify information need	Identify the user's information need	Information need
8	Obtain query terms and operators for the user's search	Obtain the initial query terms and any operators (e.g., Boolean) and apply them.	Initial query terms and operator weights
9	Retrieve and rank user query results	Provide a ranked set of relevant documents based on the submitted query	Ranked set of retrieval documents
10	Evaluate query results against information need	Assess the ranked results (Step 9) against the information need (Step 7)	Relevance judgments
11	Refine information need or query	Refine information need (Step 7) or query (Step 8) based on the evaluation findings in Step 10, such as the introduction of new terms to reduce ambiguity, removal of terms to increase diversity	Refined information need or query
12	Display final ranked set of relevant documents	Obtain and display the most correct set of relevant documents	Final ranked set of relevant retrieval documents

Table 2: Step description, task objective and expected output for the pre-processing steps of our IR model.

Step No	Step Description	Task	Expected Output
3a	Perform lexical analysis	Break each document into tokens for analysis	A set of tokens from the documents in the collection
3b	Term resolution	Resolve term acronyms and abbreviations	A set of tokens with abbreviations and acronyms resolved
3c	Tag term parts of speech	Determine and tag the part of speech for each token	Part of speech (POS) tags for each token in the collection
3d	Remove stop words	Remove specific tokens from the collection	The set of tokens determined in Step 3a, less those tokens in our stop list
3e	Stem tokens	Reduce each token to a stemmed form	A stemmed set of tokens from the collection
3f	Analyze tokens	Perform analysis on document tokens as an input into Step 5 (configure system)	An analysis of the collection, such as document statistics, diversity of tokens, etc. to determine an appropriate indexing strategy
3g	Classify documents	Assign documents to one or more classes	A set of documents contained in each class

Once preprocessing is completed, we determine the appropriate indexing strategy, configuration parameters and index the preprocessed document collection, and configure the retrieval system application. The last aspect of the system design is the creation of the search interface.

Once the system has been implemented, it is ready for searches. The right-hand portion of Figure 2 illustrates the Steps in handling queries. The user defines her or his information need and provides search terms and operators through the search interface. The interface passes this information to the search engine, which retrieves and ranks the results. The user would then evaluate the query results against her or his information need; if it was not what the user had wanted, she or he would refine the query terms and re-issue the query until satisfied with the results. The final ranked list is then displayed. These Steps are similar to Steps in traditional relevance feedback models as described in the literature, (e.g., (Baeza-Yates & Ribeiro-Neto, 1999). The description, task objective and the expected output from each step in our model appear in Tables 1 and 2.

In this thesis, we examine the Steps of this IR model, using our framework to identify those Steps most appropriate for applying crowdsourcing and/or games. We then conduct experiments on three selected Steps which have potential for the use of human computation methods. The Steps we examine are: term resolution (Step 3b), evaluate query results against information need (Step 10), obtain query terms and operators for the user's search (Step 8), and refine information need or query (Step 11).

1.2 Contributions

The contributions of this thesis are broadly as follows:

1. We introduce a framework to examine the application of human computation methods in information science with a focus on IR. This framework consists of a set of 6 criteria that allow us to evaluate the appropriateness of human computation approaches in IS. Applying this framework to our IR model, we examine if it is beneficial to use human computation methods instead of algorithmic methods for specific tasks. We also examine if it is beneficial to have human computation methods supplement current algorithm methods.

2. We identify several Steps taken from our IR model which our framework indicates that although human computation methods can be applied, no (or few) human computation methods have yet been tested in the literature. We design experiments for each of these Steps and evaluate if the obtained results support our framework. Specifically, we run experiments involving different human computation techniques, different types of agents (crowdworkers, students, algorithms) and different data collections.

3. Through the evaluation of our results applied to the selected Steps and through the use of our framework, we develop a set of guidelines for the use of human computation in information retrieval. Information science involves the analysis, collection, classification, manipulation, storage, retrieval and dissemination of information. Therefore, we extend these guidelines to new areas beyond our information retrieval system model, focusing on the applicability of human computation to other areas of information science.

4. We also discuss the limitations of our work and how human computation techniques might be affected by currently-established processes in information science.

The beneficiaries from this thesis research are those persons responsible for designing tasks that may depend on human experts to accomplish IS tasks and wish to find an alternative approach that is cheaper but maintains quality and people who currently rely on algorithms that can work quickly, but they wish to increase task quality. The implications of integrating human computation into IR tasks are that it may provide an improvement, in terms of quality or cost, on how tasks are currently being addressed. This may lead to new research in IR using human computation, yielding further improvements.

CHAPTER 2

RELATED WORK

2.1 Crowdsourcing

In his 2005 book "When Computers were Human", Grier illustrates very early uses of what is now considered crowdsourcing (Grier, 2005): in 1758, Alexis Clairault calculated the orbit of Halley's Comet by dividing the large task of numeric computations among three astronomers. Around the same time, organized computation took on the use of crowdsourcing for quality assurance when in 1760 Nevil Maskelyne calculated the moon's position for navigation by having two astronomers perform the calculations twice and a third to verify the results. Taking the idea of division of labor further, in 1794, Gaspard de Prony organized a task using ninety unemployed hairdressers – using only basic addition and subtraction -- to create detailed logarithmic and trigonometric tables. From these early examples some trends begin to appear: division of labor, mass production, and task supervisors overseeing workers with "common knowledge".

Perhaps one of the best-known examples of early crowdsourcing is the creation of the Oxford English Dictionary. Beginning in 1857 an open call was made to volunteers to index all words in the English language with example quotations for each and every one of their usages. Over the next 70 years the editors received over 6 million submissions (Simpson, 2004). Similarly, crowdsourcing has been used as a competition in order to discover the best solution to a problem. In the 19th century, the French government proposed several of these competitions created for poor Frenchmen who had performed virtuous acts. To win a prize offered during one of these open calls, Nicholas Appert invented a new way to preserve food in air-tight jars (Farkas, 2003). Similarly, the British government provided a reward to find a straightforward method to determine a ship's longitude in the Longitude Prize (Sobel, 1995). The 20th century continued this trend, introducing open call contests for creative short film

productions, radio voiceovers, and commercials jingles, all of which are viewed as precursors to what we know today as crowdsourcing.

2.1.1 Crowdsourcing Definitions and Context

Crowdsourcing was first defined by Jeff Howe in his 2006 *Wired* article (Howe, 2006) as

"the act of taking a job traditionally performed by a designated agent (usually an employee) and making it available to an undefined, generally large group of people in the form of an open call."

Crowdsourcing, which is occasionally referred to as mechanized labor, forms one of the two most popular genres of human computation (the other being games with a purpose) (Quinn & Bederson, 2011)

Crowdsourcing and outsourcing are occasionally confused with one another. *Crowdsourcing* is a distributed process where production and problem-solving tasks are outsourced to an anonymous crowd of people, while *outsourcing* involves a specific known body that agrees to assist with the process. To illustrate the difference between the two, we return to the language translation example: Outsourcing may involve a contracted agreement between a publishing house and an identified firm to translate a book into several languages. Crowdsourcing might involve multiple individual translators, each one translating a chapter of the book to a specific language. These individual translators may be graduate students studying foreign languages with some spare time or even translators for an outsourcing firm who wish to pick up some additional work on the side. One key aspect of crowdsourcing is the "open call", which allows anyone from a pool of candidates, with various levels of qualifications, to offer their services. Crowdsourcing strives to be based on meritocracy, where providing quality work is the best contributor to a worker's continued success.

There are four primary advantages of crowdsourcing: speed, cost, quality, and diversity (O. Alonso, 2011):

Speed. Tasks can be set up quickly and easily and results can often be obtained in less than 24 hours.

Low cost. Conducting tasks is usually inexpensive. A few cents per task, even with task redundancy to enhance quality, is a fraction of the compensation required if temporary in-house workers were used instead.

Respectable quality. As long as tasks are designed with the appropriate control mechanisms, results are usually of good quality (Ipeirotis, Provost, & Wang, 2010; Zaidan & Callison-Burch, 2011).

Diversity of participants. Diversity of available workers in the labor pool is good. This can be beneficial for creative tasks or when a task requires a talent you cannot easily find through standard channels (such as a translator from a rare language to English).

Figure 3 shows the relationship between different areas related to crowdsourcing.

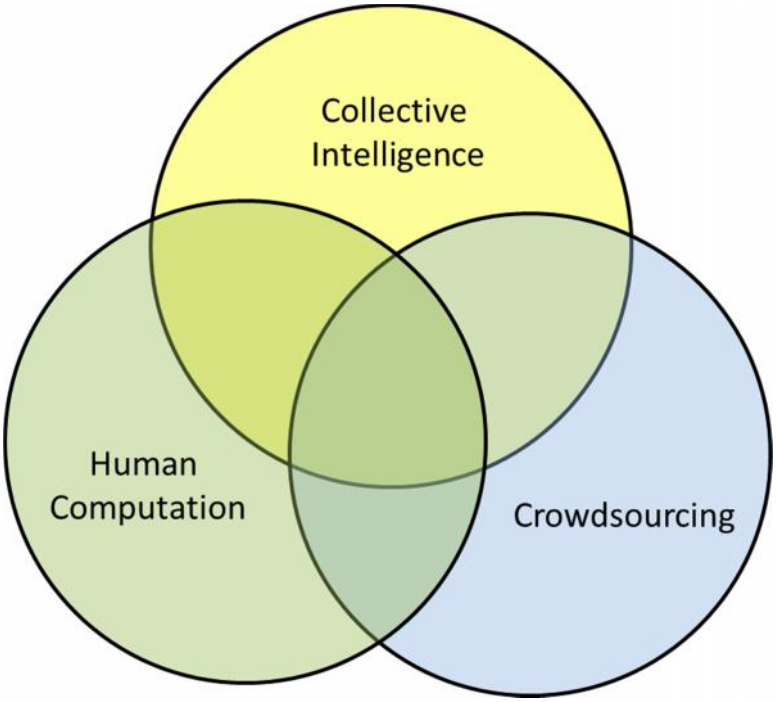

Figure 3: Relationship between human-centric computation approaches.

Human-based computation (or ***human computation***) has been defined similarly in the literature (Chan, King, & Yuen, 2009; Quinn & Bederson, 2009; Yuen, Chen, & King, 2009); we use the definition by Louis von Ahn: "a paradigm for utilizing human processing power to solve problems that computers cannot yet solve" (L. von Ahn, 2005). To this end, human computation techniques typically engage crowds, or large groups of human participants and therefore have developed into a group of techniques for harvesting collective intelligence. Additionally, since human computation focuses on problems that are not (yet) solvable by computers, it has become a useful mechanism in a number of scientific disciplines concerned with building intelligent algorithms including speech processing, Natural Language Processing (NLP), or the semantic web (Sabou, Bontcheva, Scharl, & Föls, 2013)

Human computation refers to the method in which the work is done; it replaces computers with either humans or hybrid algorithm-human models. Crowdsourcing, on the other hand, involves humans, but described how they are selected; it replaces identifiable human workers with anonymous or semi-anonymous workers hired through an open call. There is some overlap between human computation and crowdsourcing, particularly where humans and computers had previously-established roles performing the same type of task. The intersection of crowdsourcing with human computation in Figure 3 represents applications that could be considered as replacements for either traditional human roles or computer roles. For example, language translation is a task that can be accomplished either by computers when the priority is speed or cost, or by human translators when the priority is quality. A human computation approach allows the merits of both approaches to be combined; for example, the MonoTrans project (Hu, Bederson, Resnik, & Kronrod, 2011) provides a tradeoff on these three priorities with moderate speed, cost, and quality and could be considered members of both sets.

Collective intelligence is the notion that group decision making can create certain synergies that individual decision making is unable to achieve. Traditional study of collective

intelligence focused on the inherent decision making abilities of large groups (Lévy, 1999; Mataric, 1993). Malone's taxonomical "genome of collective intelligence" defines collective intelligence as "groups of individuals doing things collectively that seem intelligent." (T. W. Malone, Laubacher, & Dellarocas, 2009)

To illustrate the benefits of collective intelligence, a study on sports betting demonstrated that not only does collective intelligence increase the accuracy of determining the winner, but also reduces the variance as more people get involved (Pennock, 2007). Collective intelligence usually has a built-in feedback mechanism, which provides just-in-time knowledge for better decision-making than individuals would have acting alone (Bonabeau, 2009).

The scope of collective intelligence is broad: Malone's definition includes the Google's PageRank web indexing algorithm and any group collaboration, including "families, companies, countries, and armies." Crowdsourcing makes use of many of the principles of collective intelligence, including the "open call" nature of task requests and in some tasks the reliance on voting to provide the best solution to a given problem.

The key distinctions between collective intelligence and human computation are the same as with crowdsourcing, but with the additional distinction that collective intelligence applies only when the process depends on a group of participants. It is conceivable that there could be a human computation system with computations performed by a single worker in isolation- which is why part of human computation protrudes outside collective intelligence. To illustrate, consider the example of an individual human translator who operates an on-demand, computer-based language translation service. It is human computation because it uses the human translator's abilities to perform a computational task translating text from one language to another. It would not be considered collective intelligence because there is no group involved, and thus no "group" behavior at work.

2.1.2 Taxonomy of Crowdsourcing

Crowdsourcing can be broken down taxonomically into subcategories, as illustrated in Figure 4.

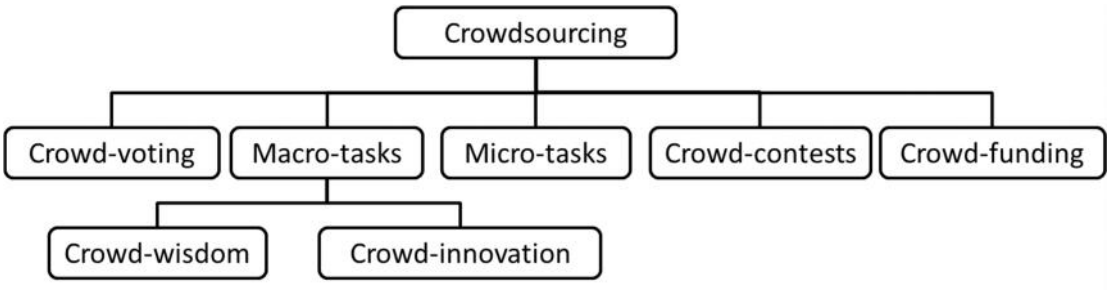

Figure 4: A taxonomy of crowdsourcing

2.1.2.1 Crowd-voting

Numerous commercial websites obtain community opinions, recommendations and creative suggestions using a process known as crowd-voting. As described in Howe's book (Howe, 2008), Threadless.com[2], a t-shirt retailer, crowdsources its entire design process by having users submit designs and then have the crowd vote on the best ones, which are then printed and sold. With tens of thousands of members who provide designs and vote on them, the website's products truly are created and selected by the crowd, rather than the company.

2.1.2.2 Macro-Tasks

Macro-tasks are large-scale projects which are presented to the crowd with a request for workers to get involved with the aspects of the task in which they are most knowledgeable. Crowd participants are empowered to follow the best course of action. Macro-tasks are suitable for Research and Development projects and for project innovation

[2] http://www.threadless.com/

ideas and are broken down further into two sub-categories: Crowd-wisdom and Crowd-innovation.

2.1.2.3 Crowd-wisdom

Crowd-wisdom collects large amounts of information from numerous sources and then aggregates the content. Wikipedia is probably the best-known example, but knowledge bases and forums established by the crowd are becoming more ubiquitous, e.g., del.ic.ious[3] (social bookmarking), Flikr[4] (photos) and YouTube[5] (user-generated video content). Hundreds of thousands of contributors participate daily; contribute their knowledge on millions of topics. In addition to the examples mentioned here, websites such as chaordix[6] and quirky[7] also make use of this format.

2.1.2.4 Crowd-innovation

Research and development provides many opportunities for crowdsourcing people's thoughts and ideas. InnoCentive[8] is a crowdsourcing platform for corporate research and development where scientific problems are posted as an open call for solutions with a cash prize that can range from $10,000 to $100,000 per challenge (Howe, 2008). Some crowd-innovation initiatives involve gathering general ideas and innovation that can be applied to a number of scenarios. These events incorporate an Internet-based platform to facilitate easy idea generation and discussion. Examples of these competitions include events like IBM's

---

[3] http://delicious.com/

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

[5] http://www.youtube.com/

[6] http://www.chaordix.com/

[7] http://www.quirky.com/

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

"Innovation Jam"[9], first held in 2006. To date, these innovation-seeking events have been attended by over 140,000 international participants and have yielded around 46,000 ideas (Salminen & Harmaakorpi, 2011).

2.1.2.5 Micro-Tasks

Micro-tasks are offered on crowdsourcing platforms where users accomplish those tasks that computers cannot perform well in exchange for small amounts of money. MTurk[10] is the best-known example of such a platform. Each task requires very little time and offers a very small amount in payment. One of MTurk's better-known projects was an open call for assistance examining satellite images to unsuccessfully locate the boat of lost researcher Jim Gray (Hellerstein & Tennenhouse, 2011). Harvard's Tuberculosis Lab teamed with Crowdflower[11], a crowdsourcing platform, to help identify drug resistant TB cells in slides showing mouse cortex images. Without the use of crowdsourcing, the project would not have had enough people evaluate the hundreds of thousands of images necessary (Bingham & Spradlin, 2011). In addition to MTurk, other websites using this format are microtask[12], clickworker[13], and lingotek[14].

2.1.2.6 Crowd-contests

Crowd-contests provide cash prizes for the single-best submission or submissions that meet an established set of guidelines. These are also called a "winner-take-all" arrangement,

---

9 https://www.collaborationjam.com/

10 https://www.mturk.com/

11 http://crowdflower.com/

12 http://www.microtask.com/

13 http://www.clickworker.com/en/

14 http://www.lingotek.com/

where the crowd participants submit work based on a stated objective in an attempt to win a prize. One (or a few) winners are selected from all received submissions and are compensated; all unselected submissions receive no compensation. This can be considered an "all-pay" arrangement since the crowd pays with their labor and receives no guarantee of compensation. Chinese crowdsourcing platforms called *witkeys*, frequently use this arrangement (Liu, Yang, Adamic, & Chen, 2011; Yang, Adamic, & Ackerman, 2008) – taskcn[15] and zhubaijie[16] are two examples. Websites such as 99designs[17], crowdspring[18], and squadhelp[19] use this format. In his 2013 paper, Araujo investigated the quality of 38,000 design contests on the website 99designs and found that although designers participate, most contests are won disproportionately by a few participants (Araujo, 2013). This backs up findings from others on other "winner-take-all" crowdsourcing websites such as taskcn (Liu, *et. al.*, 2011).

Crowd-contests such as the Dorito's popular "Crash the Superbowl" advertising campaign, which solicits crowd entries for use in their Superbowl television advertisements, also uses this format. Another well-known example of a crowd-contest was the 2009 DARPA experiment, where DARPA placed 10 balloon markers in various locations across the United States and challenged teams to compete to be the first to report the location of all the balloons. A team from MIT was able to locate all ten balloons in nine hours, winning a $40,000 prize ("Darpa Network Challenge," 2009).

---

15 http://www.taskcn.com/ (in Chinese)

16 http://www.zhubajie.com/ (in Chinese)

17 http://99designs.com/

18 http://www.crowdspring.com/

19 http://www.squadhelp.com/

2.1.2.7 Crowd-funding

With crowd-funding, the crowd is asked to donate a defined amount of money for a specific cause, project or other use within a specified timeframe. If the goal is not met, the donated money is refunded. This arrangement is used for funding disaster relief, project fundraising, market research, artistic support and seeding startup activity. Websites such as crowdrise[20], kickstarter [21]and seedups[22] use this format.

2.1.3 Crowdsourcing Techniques

A variety of techniques are used for crowdsourcing tasks. These facilitate the effective gathering of information and can be divided into two groups: explicit and implicit.

Explicit crowdsourcing encourages workers to collaborate in order to evaluate, share, and complete specific tasks. For example, users may be asked to evaluate books or webpages, share user-generated content (UGC) on websites such as YouTube. Users can be asked to build artifacts, such as Wikipedia. Explicit crowdsourcing can be applied to the micro-task market, where small tasks can be selected and completed by workers for a small monetary reward. The most popular example of an explicit crowdsourcing platform is MTurk. With explicit crowdsourcing, accomplishing the tasks provided to the crowd is the primary goal.

Implicit crowdsourcing embeds the crowdsourcing task as one component used to achieving another, more important objective. It involves one of two technique subtypes: standalone and piggyback. *Standalone* techniques allow problem solving to occur as a side effect of a primary task, whereas *piggyback* techniques make use of user information and activities on third-party websites. In terms of standalone implicit crowdsourcing, an example

---

[20] http://www.crowdrise.com/

[21] http://www.kickstarter.com/

[22] http://www.seedups.com/

is the ESP Game(Luis von Ahn & Dabbish, 2004), where users clarify what an image contains and then generate a label to describe the image contents. The reCAPTCHA task asks people to solve CAPTCHAs in an effort to "prove" they are human, while providing CAPTCHAs from old books that cannot be deciphered by computers in order to try and digitize them for the web (L. Von Ahn, Blum, Hopper, & Langford, 2003). reCAPTCHA is available as a Web service by companies which need to protect themselves from spammers; the human not only authenticates himself or herself, but also contributes to the task of digitizing books (the crowdsourcing task in focus). As of late 2010, two hundred thousand sites are using reCAPTCHA, resulting in 85 million words recognized per day, which is roughly equivalent to about two million books per year (Luis von Ahn, 2013). The reCAPTCHA task takes about the same amount of time as a CAPTCHA, as it is easier to type recognized words than random characters. In the reCAPTCHA example, the primary objective is to ensure access is granted to humans only, but a side task is the deciphering of illegible text from old books.

Piggyback crowdsourcing is more subtle. It is used most frequently by companies such as Google that mine one's search history and websites in order to discover keywords for ads, spelling corrections, and locating synonyms. Thus, users in the crowd are assisting the modification and refinement of existing systems, such as Google's ad words, but the crowd assistance serves as only an input to the primary task. Thus, with implicit crowdsourcing, the tasks given to the crowd are only a single component to facilitate a larger set of goals.

2.1.4 Use of Crowdsourcing for Problem Solving

As mentioned earlier, despite the rapid growth of crowdsourcing and game mechanisms, few attempts have been made to classify the types of tasks these mechanisms address. A 2011 study by Quinn and Bederson classified human computation systems, including crowdsourcing (Quinn & Bederson, 2011) but did not describe the tasks these systems address in detail. Yuen *et. al.* perform a more comprehensive survey on tasks in

crowdsourcing (Yuen, King, & Leung, 2011) and a similar survey of serious games (Yuen, *et. al.*, 2009). In each survey, the authors provide a useful taxonomical classification, but these authors do not discuss the applicability of each aspect of their classification to real-world scenarios, such as to IR systems.

Lease and Alonso also conducted several workshops that address specific issues with crowdsourcing, games, and IR, such as addressing spam, incentive models, and quality control, among others (O. Alonso & Lease, 2011a, 2011b; Lease & Yilmaz, 2012). The Human Computation workshops examine issues in human computation, a broader classification of human-algorithm collaboration that includes both games and crowdsourcing (P. Ipeirotis, R. Chandrasekar, & P. Bennett, 2010; P. G. Ipeirotis, R. Chandrasekar, & P. Bennett, 2010). The topics covered in these workshops, although broad, provide useful insight that we employ in our research methodology.

Furthermore, as mentioned in the introduction, IR models have been discussed extensively in the literature (e.g.,Baeza-Yates & Ribeiro-Neto, 1999; Croft, Metzler, & Strohman, 2010; Manning, Raghavan, & Schutze, 2008); however, none of these discussions directly address how human computation mechanisms can add value to the Steps in our IR model.

2.1.5 Using Crowdsourcing in Information Retrieval

There have been several recent tutorials on social computing that examine the application of crowdsourcing to areas of IR. Alonso has presented some guidelines for conducting studies using crowdsourcing platforms (Omar Alonso & Ricardo Baeza-Yates, 2011; O. Alonso & Lease, 2011a). Likewise, Ipeirotis (Ipeirotis & Paritosh, 2011) has provided some useful insight into study design. However, these discussions primarily focus on crowdsourcing studies in the most high-demand tasks: labeling, translations and transcriptions, relevance judgments, and classification, ignoring many other tasks, as listed in

Tables 1 and 2. The last two, relevance judgments, and classification, are elements of our model and their work is incorporated into our research design.

To date, most crowdsourcing studies in IR have examined relevance assessment. Several studies, such as (Omar Alonso & Mizzaro, 2012; McKibbon *et. al.*, 1990) have compared the crowd to experts in document assessment, concluding there is little difference in quality, particularly when multiple assessors are used. A few evaluations have been conducted to compare crowd-based and lab-based participants on search performance. In prior work we compared crowd and lab participants on multimedia search results in (Harris, 2012), concluding that the two groups were indistinguishable in quality.

Integrating the crowd is becoming more commonplace for difficult searches, perhaps indicating that the crowd represents a nice tradeoff between speed, cost, and quality. Bozzon *et. al.* describe a tool called CrowdSearcher, which utilizes the crowd for difficult searches (Bozzon, Brambilla, & Ceri, 2012). A study by Yan *et. al.* describe a mobile search application in (Yan, Kumar, & Ganesan, 2010); claiming a search precision of 95%. Ageev *et. al.* conducted an experiment to evaluate crowd search techniques in (Ageev, Guo, Lagun, & Agichtein, 2011), but do not compare the crowd's performance with other groups. These studies support the premise that the crowd can be used to search effectively and deliver results with reasonable precision. Few evaluations have been conducted to compare crowd-based and lab-based participants on search performance. One study compared crowd and lab participants on multimedia search results in (Harris, 2012), concluding that the two groups were indistinguishable in quality.

Some of the challenges with integrating crowdsourcing into an IR system stem from the time necessary to complete certain types of tasks, as finding qualified participants in the crowd at the time of need may not always be feasible. One possible solution is to have a portion of the crowd "on-call" at a set minimum fee. This design is suggested in Yan *et. al.*'s CrowdSearch, a crowd-based image search system that discusses such an on-call system to reduce the latency time (Yan, *et. al.*, 2010).

2.2 Games

A recent research study indicated there are more than half a billion people worldwide playing online games at least an hour a day, 183 million in the US alone (McGonigal, 2011). Other studies have estimated that the average American has played 10,000 hours of video games by the age of 21, equivalent to five years of working a full-time job 40 hours per week. (Richards). What if some of this time and energy could somehow be channeled into productive work? And better yet, what if people playing computer games could, without consciously doing so, simultaneously solve large-scale computation problems?

This redirection of a user's time and energy from pure entertainment to accomplishing a task while being entertained is a principle factor behind the development and use of games with a purpose. Other factors include the ability to make mundane tasks engaging, lower cost-per-task than even crowdsourcing might claim, and a reduction in spam over crowdsourcing inputs. Although computers have made phenomenal advances over the past half century, they still do not possess the conceptual intelligence or perceptual capabilities most humans take for granted. If human brainpower is treated as part of a large distributed network, each node is able perform a tiny fragment of a large computation. The trouble is, unlike computer processors, humans need to be given an inducement or incentive to become part of a collective computation. Thus, the allure of online games works as a seductive method for encouraging people to participate in this process.

2.2.1 History of Serious Games

Military officers have been using scenario-based war games in order to train strategic skills for centuries. This may be the first connection to what we consider serious games today. In his 1970 book, Clark Abt discussed the idea and coined the term "serious games" (Abt, 1987). Abt's examples referred to board and card games, but he gave a useful general definition which is still considered applicable today:

"Reduced to its formal essence, a game is an activity among two or more independent decision-makers seeking to achieve their objectives in some limiting context. A more conventional definition would say that a game is a context with rules among adversaries trying to win objectives. We are concerned with serious games in the sense that these games have an explicit and carefully thought-out educational purpose and are not intended to be played primarily for amusement."

The modern concept of "Human Computation Game" or "Social Game" was first coined in 2006 by Luis von Ahn *et. al.,* who created games with a purpose (L. Von Ahn, 2006). These games produced useful data as a by-product. They solve some problems that computers cannot currently resolve, while at the same time, take advantage of people's desire to be entertained.

Although serious games are designed to be entertaining, their main purpose usually falls within one of three categories: to train, to investigate, or to advertise (Groh, 2012). They are denoted "serious games" to differentiate them from games designed only for leisure. A subset of serious games, designed to be played by the crowd through an open call, is denoted games with a purpose (GWAP). In this thesis, we refer to "GWAP" as "serious games" or simply, "games".

2.2.2 Serious Games Definitions and Context

Serious games are human-based computation techniques designed for a primary purpose other than pure entertainment, e.g., to solve a specific problem. ***Games with a Purpose (GWAP)*** are a type of serious game in which a computational process performs its function by crowdsourcing certain Steps to humans in an entertaining way (L. Von Ahn, 2006). Thus a GWAP can be considered the intersection between serious games and crowdsourcing. The approach most commonly taken by GWAPs is to exploit the differences in abilities and alternative costs between humans and algorithms in order to achieve symbiotic human-computer interaction. The game's true purpose may either be stated or kept

hidden from the player. GWAPs have a vast range of applications in areas as diverse as security, computer vision, adult content filtering, and Internet search.

Gamification is the use of game design techniques and mechanics to enhance non-games to improve user experience and user engagement (Deterding, Sicart, Nacke, O'Hara, & Dixon, 2011). Gamification makes technology more engaging, by encouraging users to engage in desired behaviors, and by taking advantage of the human psychological predisposition to engage in gaming (Thomas W Malone, 1980, 1982; Radoff, 2011). With gamificaton, humans are encouraged to perform tasks ordinarily considered mundane, such as completing user surveys, annotating images, or rating and reviewing product web sites. Huotari and Hamari indicate that gamification invokes the same psychological experiences as games generally do (Huotari & Hamari, 2012).

One reason gamification has received considerable attention in recent years is for its ability to enhance positive patterns of service use, such as increasing user activity, social interaction, quality and productivity (Hamari, 2013). These actions are considered to be a result of positive, intrinsically motivating, 'gameful' experiences (Huotari & Hamari, 2012). It is reflected in an academic context as well, with a growing number of papers that discuss the topic of gamification. The growth of the term 'gamification' in paper titles has increased rapidly as well, suggesting that interest in gamification as a subject for academic inquiry is increasing (Hamari, Koivisto, & Sarsa). To minimize jargon and enhance readability, in this paper, we will refer to gamification as 'applying game-based mechanisms.'

This new gamification trend has already gained some traction. Gartner has estimated that over 50 percent of organizations managing innovation processes will incorporate game-based mechanisms in their business by 2015 (Gartner, 2011). Additionally, an increasing number of successful startups have organized their offerings around adding a game-based mechanism to their existing services, such as Codeacademy[23], a service that uses game-like

[23] http://www/codeacademy.com

elements to help teach users how to code or Badgeville[24], which offers a software-as-a-service (SaaS) based technology for web and mobile sites to measure and influence user behavior through game-based techniques.

We follow Yuen *et. al.*'s categorization of serious games using four features: game structure, verification method, game mechanism and required number of players (Yuen, *et. al.*, 2009). Table 3, taken from Yuen, *et. al.*, 2009, illustrates this categorization.

Table 3: Categorization of serious games

Game Structure	**Verification Method**	**Game Mechanism**
Output-agreement	Symmetric	Collaborative or Hybrid
Input-agreement	Symmetric	Collaborative or Hybrid
Inversion-problem	Asymmetric	Collaborative, Competitive or Hybrid
Output-optimization	Asymmetric or Symmetric	Collaborative, Competitive or Hybrid

Game Structure defines the key elements of a game including player inputs and outputs, the relationship among these inputs and outputs for all players, and the winning condition. In (L. Von Ahn & Dabbish, 2008), von Ahn defined three types of games: output-agreement games, input-agreement games and the inversion-problem games. Some of the more recent games cannot be categorized into von Ahn's three categories, such as the Restaurant Game and Diplomacy, which aims to collect player output patterns or behaviors. Yuen *et. al.* defines a new category, called an output-optimization game, for this last game category.

In *output-agreement games*, each player is provided with the same inputs (e.g. the same image) and the objective is to produce outputs (e.g. annotations of that image) based on

[24] http://www.badgeville.com

the common input. With the ESP Game[25], where users try to label images, two players paired at random are provided an image and a set of forbidden, or "taboo", words that cannot be used as labels. The player's objective is to provide matching labels for the same image; if the players enter the same label, it becomes metadata for the image and a "taboo" word for future rounds. In *input-agreement games*, each player is provided information in the game that may be either the same or different. The players are instructed to produce outputs describing their input with the objective of assessing whether player inputs are the same or different. Players are only able to see each other's outputs. With TagATune[26], two users paired at random are provided with music and have to describe it by typing labels. The two players see what each other typed and must determine if they are listening to the same tune. In *inversion-problem games,* the first player has access to the whole problem and is responsible to provide hints to the second player to help guess the answer to a puzzle. If the second player is able to guess correctly, it is assumed that that the hints given by the first player are correct and meaningful clues. In Verbocity[27] two players paired at random take turns with two different roles: describer and guesser. As describer, you help the other player guess a secret word by giving clues. As guesser, you use these clues to guess a secret word. This information is used to evaluate the utility of the clue to providing the secret word. Last, in *output-optimization games,* each player is provided with the same input and their outputs are hints of other players' outputs. With Diplomacy[28], a board game with the objective of controlling a number of supply centers, users are given the same inputs and must make a determination of how to react based on the moves of adjacent players. The information

---

25 http://www.gwap.com/gwap/gamesPreview/espgame/

26 http://www.gwap.com/gwap/gamesPreview/tagatune/

27 http://www.gwap.com/gwap/gamesPreview/verbosity/

28 http://www.playdiplomacy.com

gained from this game is used to evaluate negotiation strategies in autonomous agent algorithms.

Verification Methods define the method used to check player output accuracy by asking players to perform either the same task or a different task in a game. In *symmetric verification games*, each player is asked to perform the same task and their outputs are compared with one another. The ESP game is an example of a symmetric verification game. In *asymmetric verification games*, each player is asked to perform a different task (e.g., in the asymmetric verification game Verbocity, one player, called the "narrator", provides clues to a certain word; the other player, called the "guesser", tries to deduce the word that the narrator's clues describe).

Game Mechanisms define the relationships between game participants. In *collaborative games*, for all players to win as a team, each player must complete the task assigned to her or him, which in turn helps other players to complete their tasks. The ESP game is an example of a collaborative game. In *competitive games*, a player has to complete the task assigned to her or him. The player's achievement is compared with the achievements of other players, the player's previous game history, or information stored in a database. PageHunt[29] is a game where the player is served random pages from the Internet. The player is tasked with developing a query that would bring up this page in the top few results on a search engine. The web page being 'hunted' is shown in the background. The player types in queries and looks at results returned in the floating operating panel, initially in the lower right corner. This is scored and the score is evaluated in comparison with other players. In *hybrid games*, players have to complete their assigned tasks, which aid other players to complete their tasks. Each player's achievement is then compared with the achievement of other players, the player's previous game history, or information stored in a database. In the

[29] http://pagehunt.msrlivelabs.com/PlayPageHunt.aspx (link discontinued as of Oct 2011)

hybrid game, Phetch[30], three to five players label images on the Internet with descriptive captions. The purpose of this game is to assist sight-impaired readers. One player is designated as a describer, while the others are seekers. The describer is shown an image, which she or he uses words to describe to the seekers. The seekers use an Internet image search engine to attempt to find the image being described. The first seeker to find the image gains points and becomes the describer for the next round. The describer is also rewarded for a successful outcome.

Player Requirements define the rules of number of players and whether players play against each other concurrently or consecutively. In *synchronous games*, players access the same game at the same time and can provide a real-time response to other players' actions. The ESP Game and TagATune are synchronous games. In *asynchronous games*, players do not provide a real-time response to other players' actions. The information collected from one player is stored in a database and is used to determine the correctness of the output provided by other players. PageMatch, a variation of the PageHunt game where players compete by coming up with queries for a given webpage in the least amount of time, is an asynchronous game; players can play PageMatch at different times and compete by achieving the highest score. *Single-player* games allow one player to participate in a game. Moves of other players can be simulated from a pre-recorded game. PageHunt is a single player game. *Two-player games* allow two players to participate in a game. The ESP Game and Verbocity are two-player games. Likewise, *multi-player games* do not define an upper limit to the number of players able to participate in a specific game.

In GeAnn[31] (Games for Engaging Annotations) (C. Eickhoff, Harris, Srinivasan, & de Vries, 2011; Lease & Yilmaz, 2012), players are given several categories in a Tetris-like

---

[30] http://www.peekaboom.org/phetch (link discontinued)

[31] http://www.geann.org

game to describe how a given term matches a given portion of text, and are subsequently rewarded by correctly matching the consensus category of other players. GeAnn is designed for obtaining relevance judgments when there is no appropriate gold standard against which to compare. Using the above framework, GeAnn would have the following characteristics. It is an output-agreement game, since players are provided with the same inputs – a block of text, a keyword, and a few categories that describe the keyword's potential association with the text block – and the players attempt to provide the majority decision category for that keyword. It is symmetric-verification, since players perform the same task and outputs are compared with those from other players. It is a hybrid game mechanism, since players must complete a given task, and the outputs from this task enable other players to accomplish their task in the form of majority decisions on the keyword-text association. It is an asynchronous single-player game, since one player competes at a time, and the players do not provide a real-time response to the actions of other players.

Table 4, derived from Yuen, *et. al.*, 2009, provides an overview of this categorization and a number of serious games that exist in each category. For areas marked 'N/A', the features are certainly possible, but no game has been observed for that feature set.

2.2.3 Use of Serious Games in IR

A number of the games mentioned here involve some aspect of Information Retrieval (IR); however, most involve tasks considered supplemental to IR, such as image labeling or genre classification. Only a few address core IR tasks directly, particularly in search engine evaluation. Unlike crowdsourcing, which usually involves a single task that accomplishes several complementary objectives at once, serious games usually focus on a single objective.

Table 4: Examples of serious games and their features.

Game Structure	Verification Method	Game Mechanism	Player Requirements		Examples
			No. of Players	Game Play	
Output-agreement	Symmetric	Collaborative	2	Synchronous	ESP*, Match*, Squigl*, OntoGame, Page Match*
		Competitive	1 or 2	Asynchronous	Page Hunt*, Page Race* FuFinder, Book Explorer
		Hybrid	Multi	Synchronous	Common Consensus*, Social Heroes*
		Hybrid	Multi	Asynchronous	Gopher Game*, GeAnn*
Input-agreement	Symmetric	Collaborative	2	Synchronous	TagATune*
		Hybrid	N/A	N/A	N/A
Inversion-problem	Asymmetric	Collaborative	1 or 2	Synchronous	Peekaboom*, Verbocity* KissKissBan*
		Competitive	2	Asynchronous	Dogear, CyPress CARS
		Hybrid	1, 2 or Multi-player	Synchronous	Phetch*
Output-optimization	Symmetric	Collaborative	2	Synchronous	Restaurant Game
		Competitive	N/A	N/A	N/A
		Hybrid	Multi-player	Synchronous	Diplomacy
	Asymmetric	Collaborative	N/A	N/A	N/A
		Competitive	N/A	N/A	N/A
		Hybrid	N/A	N/A	N/A

The ESP Game (Luis von Ahn & Dabbish, 2004) is considered the first game-based computation system. It has been subsequently adopted as the Google Image Labeler. Its

objective is to collect image labels for Google Images. In addition to basic image annotation, Peekaboom (L. Von Ahn, Liu, & Blum, 2006) is designed to have players determine object locations within images. Squigl (Łupkowski, 2011) provides outlines for each object within an image. Phetch (L. von Ahn, Ginosar, Kedia, & Blum, 2007; L. Von Ahn, Ginosar, Kedia, Liu, & Blum, 2006) is designed to enhance image descriptions and therefore improve web accessibility, particularly with image searches. Likewise, to provide a tool to help visually-impaired people navigate the internet, WebInSight (Bigham, Kaminsky, Ladner, Danielsson, & Hempton, 2006) was designed to automatically create alternative text in documents when it is lacking in a document. Using a game format based on the ESP Game and Phetch, WebInSight generates a collection of alternative text. In a similar way, Matchin (Hacker & Von Ahn, 2009) helps search engines rank images that fit a given set of criteria. These games improve the inputs to IR, and could be akin to the preprocessing Step for non-text retrieval.

The underlying concept of the ESP Game has been applied to other aspects of Information Retrieval. For instance, TagATune (E. Law & Von Ahn, 2009) provides annotation for sounds and music, which in turn can improve searches for audio clips. Verbosity (L. Von Ahn, Kedia, & Blum, 2006) and Common Consensus (Lieberman, Smith, & Teeters, 2007) collect "common-sense" knowledge. This common-sense information can be provided as inputs to situations involving reasoning and is a Step towards enhancing interactive user interface design.

Recently, several geospatial tagging games have been introduced. Like the image labeling games discussed previously, these also focus on creating geo-tagged inputs to IR. MobiMission (Grant *et. al.*, 2007) is a location-based pervasive social game played using mobile devices in which "missions" are created, enacted, and reviewed by players. MobiMission attempts to assign missions near players (e.g., find the Thai restaurant closest to Seattle's Space Needle). If there are no nearby missions available, it assigns location-independent missions instead (e.g., take a picture of a Thai restaurant with a takeout menu in

your area). The Gopher game (Casey, Kirman, & Rowland, 2007) takes a similar approach: missions are created, enacted, and reviewed by players, except that the system limits missions to players to a certain vicinity. CityExplorer (Matyas, 2007; Matyas *et. al.*, 2008) is a location-based variant of the popular German board game Carcassonne[32]. It only permits players to place tokens (markers) in predefined types of real-world locations, so players are forced to explore the unstructured game area, and the system collects new geospatial data as a by-product. Moreover, Eyespy (Bell *et. al.*, 2009) allows players to tag geographic locations using photos or text or confirm the locations of places that other players have tagged. As a result, Eyespy produces a collection of recognizable and locatable markers as well as geographic details for location-based applications.

More directly relevant to IR, Page Hunt (Ma, Chandrasekar, Quirk, & Gupta, 2009) is a single-player game that displays a random web page to a player; the player's task is to come up with query terms and operators that would bring up the web page in the top few results (e.g., the top 5) of a search engine. PageHunt evaluates a player's use of appropriate query terms and operators, an essential Step in IR. Page Race and Page Match are two variations of the Page Hunt game that are two-player competitive and collaborative games, respectively. With Search War (Edith Law, Ahn, & Mitchell, 2009), players provide an evaluation of a search page's relevance to a particular search query as well as its most salient purpose. A pair of players is each given a unique search query with the goal of guessing their opponent's search query first. To accomplish this goal, players must choose among a set of webpages associated with their search query to show their partner, such that their partner will have difficulty guessing. The premise is that players will select the worst webpage, and the game produces additional search terms. Thus Search War provides tests the ability for

[32] http://jcloisterzone.com/en/

players to create and refine queries, as appropriate, to evaluate their partner's unstated information need.

To ensure the quality of the collected labels, some serious game implementations adopt consensus opinions as the correctness measure. Picture This (Bennett, Chickering, & Mityagin, 2009a, 2009b) is a game designed to elicit relative relevance judgments from users to rank images with respect to an image query. With KissKissBan (Ho, Chang, Lee, Hsu, & Chen, 2009), image labeling is improved by involving three players, with one player attempting to prevent two players from collaborating. Therefore, it has both collaborative and competitive aspects, but it is still a game designed to improve IR inputs, not outputs.

From our discussion so far, it may appear many IR tasks can be accomplished equally through either serious games or crowdsourcing mechanisms. However, each mechanism has its own specific merits and requirements. Crowdsourcing is typically more suitable for micro-tasks that are short in duration, more variable in format (such as solutions of InnoCentive problems), or where capturing worker interest is not as essential. In contrast, serious games are typically more suitable for tasks which are larger in scale or longer in duration, such as image labeling campaigns, due to the increased overhead required to create them. Also, games are better suited for tasks with a standardized input or are focused on one specific objective, such as classifying a set of photos in certain categories. Additionally, games are more appropriate for mundane tasks such as the collection of more unique image labels for images. Finally, they are important for tasks that require a higher level of user engagement. Although usually more costly to set up, games have been shown to be cheaper in the long run for mundane tasks such as relevance judgments (C. Eickhoff, *et. al.*, 2011).

2.3 Comparison of the Advantages and Disadvantages of Crowdsourcing and Games

In the literature to date, there have been few studies that examine the use of crowdsourcing and serious games under similar conditions. Those that rely on literature

review identify some differences between these two human computation genres, but do not provide indications of the overlap or range of differences (e.g., exactly how much faster is the use of crowdsourcing compared with use of a serious game when used in the same context?) Sabou *et. al.* (2013) examined a number of studies that compared the two; a subset of their results across the features that would potentially relate to IR tasks are summarized in Table 5.

Examining Table 5 in more detail, we see there are some clear distinctions between the two genres across features, but also some ambiguous findings. For cost, the higher setup costs of designing a GWAP are offset by the lower price paid per task. The set-up time for crowdsourcing tasks as well as the higher throughput (the amount of data created per human hour) make crowdsourcing tasks more advantageous from the perspective of task speed. Thaler *et. al.* indicate that the throughput time necessary for their OntoPronto game was twice that needed for crowdsourcing a comparable task (Thaler, Simperl, & Wolger, 2012). Likewise, Chamberlain report average throughput of 550 per hour for their GWAPs, which compares unfavorably to the near-instant completion time of crowdsourcing (Chamberlain, Fort, Kruschwitz, Lafourcase, & Poesio, 2013).

For quality-based features, the balance is more evident. Chamberlain, *et. al.*, (2013) and Wang *et. al.* (2013) have independently determined that better quality could be found through games. On the other hand, Thaler (Thaler, *et. al.*, 2012) found that the quality was similar between the two for their game. In our own experiments with GeAnn, we found the quality of the GWAP compared favorably to crowdsourcing based tasks, although our method combined the two genres –we initially recruited participants by offering them a small sum to participate, but encouraged them to continue without additional compensation (C. Eickhoff, Harris, de Vries, & Srinivasan, 2012).

Table 5: Advantages and disadvantages of crowdsourcing and GWAP platforms, as applied to potential IR tasks

Feature	Crowdsourcing	GWAP	Related References
Cost			
Set-up Price	Low (+)	High (-)	(Poesio, *et. al.*, 2013; Thaler, *et. al.*, 2012; Wang, Hoang, & Kan, 2013)
Price per Task	Low (-)	None (+)	(Poesio, *et. al.*, 2013; Thaler, *et. al.*, 2012)
Speed			
Set-up Time	Low (+)	High (-)	(Poesio, *et. al.*, 2013; Thaler, *et. al.*, 2012; Wang, *et. al.*, 2013)
Throughput	High (+)	Low (-)	(Chamberlain, *et. al.*, 2013)
Quality			
Quality	Low (-)	High (+)	(Chamberlain, *et. al.*, 2013; Wang, *et. al.*, 2013)
	High (+)	High (+)	(Thaler, *et. al.*, 2012)
Maintaining Motivation	Easy (+)	Difficult (-)	(Thaler, *et. al.*, 2012)
Incentive to Cheat	High (-)	Low(er) (+)	(Chamberlain, *et. al.*, 2013; Wang, *et. al.*, 2013)
Importance of Task Interestingness	Low (+)	High (-)	(Thaler, *et. al.*, 2012; Wolf, Knuth, Osterhoff, & Sack, 2011)
Worker Diversity	Low (-)	High (+)	(Thaler, *et. al.*, 2012)
	High (+)	Low (-)	(Wang, *et. al.*, 2013)

For quality-based features, the balance is more evident. Chamberlain, *et. al.*, (2013) and Wang *et. al.* (2013) have independently determined that better quality could be found through games. On the other hand, Thaler *et. al.* found that the quality was similar between the two for their game. In our own experiments with GeAnn, we found the quality of the GWAP compared favorably to crowdsourcing based tasks, although our method combined the two genres – we initially recruited participants by offering them a small sum to

participate, but encouraged them to continue without additional compensation (C. Eickhoff, *et. al.*, 2012).

Since crowdsourcing participants are motivated by financial compensation, the ability to maintain motivation is easier than for GWAP. However, the incentive to cheat in a GWAP is lower, since there is usually no form of extrinsic compensation provided. Similarly, maintaining task interestingness is less important in crowdsourcing for this very reason - participants are motivated by compensation and will accept tasks independent of their level of interestingness.

It is unclear if worker diversity is reduced or enhanced by either crowdsourcing or GWAPs. Statistics have shown that a small number of people participate in a large number of crowdsourcing tasks (Fort, Adda, & Cohen, 2011; Poesio, Chamberlain, Kruschwitz, Robaldo, & Ducceschi, 2013). Thaler found that games reached a larger player base than crowdsourcing task workers (Thaler, *et. al.*, 2012), but others found that games may provide a larger variety of contributors and can reach more individuals across a wider spectrum than MTurk (Parent & Eskenazi, 2011) .

2.4 Integration of Crowdsourcing and Games into IR

In summary, the application of crowdsourcing has been concentrated on a few IR tasks. To illustrate this point, consider the IR model (Figure 2): approximately three dozen separate studies have investigated the application of crowdsourcing to those Steps of the IR model where algorithms performance could be improved, including eight studies on obtaining document collections (Step 2) and another ten studies that evaluate query results against an information need (Step 10). In addition, there are six research studies that use crowdsourcing to identify an information need (Step 7) and four studies that use the crowd to obtain query terms and operators (Step 8). Last, some Steps of the IR model have little, if any, representation in the literature, such as term resolution (Step 3b), stem tokens (Step 3e) and the analysis of tokens (Step 3f).

Regarding the application of games to IR, the existing research is narrower in scope than with crowdsourcing, with fewer examples. As with crowdsourcing, evaluating query results against an information need (Step 10) is well-represented in the literature with approximately a dozen studies, but only two studies that examine the use of games to obtain a document collection (Step 2). There are three studies in the literature that use games to identify an information need (Step 7) and four that use games to obtaining query terms and operators (Step 8). Tables 6 and 7 illustrate some of the existing applications of crowdsourcing and games, respectively, to each step of our model.

Table 6: Focus of literature on of crowdsourcing and games in IR

Step No	Step Description	Task Objective	Recent research and work in Crowdsourcing	Recent research and work in Games
1 (SD)	Define domain	Define the scope of information for retrieval tasks	Various areas, such as pharmaceuticals(Ekins & Williams, 2010) and general knowledge (Baumeister, Reutelshoefer, & Puppe, 2007).	No relevant games found.
2 (SD)	Obtain document collection	Determine the documents to be indexed and available for retrieval. Web spidering, web link analysis, etc., could be done at this stage.	Web collaboration and community knowledge systems: (Baumeister, *et. al.*, 2007; Chklovski & Gil, 2005; Lawrence, 2011; McCann, Doan, Varadaran, Kramnik, & Zhai, 2003; R. M. C. McCreadie, Macdonald, & Ounis, 2010; Richardson & Domingos, 2003a, 2003b; Simon, Haslhofer, & Jung, 2011); User submitted bookmarks: (Parry)	Verbosity (L. Von Ahn, Kedia, *et. al.*, 2006) and Common Consensus (Lieberman, *et. al.*, 2007)
3 (SD)	Preprocess documents	Provide data enrichment	See Table 7 for more details on each step	See Table 7 for more details on each step
4 (SYS)	Index documents	Create an index for all documents	No relevant crowdsourcing work found	No relevant games found.
5 (SD)	Configure retrieval system	Determine the best search strategies and parameters for an anticipated set of tasks	No relevant crowdsourcing work found	No relevant games found.
6 (SD)	Implement user interface	Have an interface that users can use to meet their needs	Mozilla open-source project	No relevant games found.
7 (U)	Identify information need	Identify the user's information need	Collaborative approaches: (Odumuyiwa & David, 2009; Zuccon, Leelanupab, Whiting, Jose, & Azzopardi, 2011).	Find It If You Can (Ageev, *et. al.*, 2011); Page Hunt (Ma, *et. al.*, 2009); Search War (Edith Law, *et. al.*, 2009)

Table 6 Continued

Step No	Step Description	Task Objective	Recent research and work in Crowdsourcing	Recent research and work in Games
8 (U)	Obtain query terms and operators for the user's search	Obtain the initial query terms and any operators (e.g., Boolean) and apply them.	Several studies have evaluated query terms, but only peripherally, such as Ageev (Ageev, *et. al.*, 2011)Brenes (Brenes *et. al.*, 2009); Grady(Grady & Lease, 2010); Snow (Snow, O'Connor, Jurafsky, & Ng, 2008)	Find It If You Can (Ageev, *et. al.*, 2011); Page Hunt (Ma, *et. al.*, 2009); Search War (Edith Law, *et. al.*, 2009); Koru (Milne, Nichols, & Witten, 2008);
9 (SYS)	Retrieve and rank user query results	Provide a ranked set of relevant documents based on the submitted query	No relevant crowdsourcing work found	No relevant games found
10 (U)	Evaluate query results against information need	Assess the ranked results (Step 9) against the information need (Step 7)	Many papers cover using crowdsourcing for relevance assessment, including Alonso (O. Alonso & R. Baeza-Yates, 2011; O. Alonso & Mizzaro, 2009), Blanco (Blanco *et. al.*, 2011), Grady (Grady & Lease, 2010), Nowak (Stefanie Nowak & Ruger, 2010; S. Nowak & Rüger, 2010), Brenes (Brenes, *et. al.*, 2009); Macreadie (R. M. C. McCreadie, *et. al.*, 2010); Whitehill (Whitehill, Ruvolo, Wu, Bergsma, & Movellan, 2009) Many, including ranking twitter feeds: (Naveed, Gottron, Kunegis, & Alhadi, 2011); Music: (Urbano, Morato, Marrero, & Martín, 2010) Rankr: (Luon, Aperjis, & Huberman), Learning to Rank methods (Kumar & Lease, 2011), etc.	Find It If You Can (Ageev, *et. al.*, 2011); Page Hunt (Ma, *et. al.*, 2009); GeAnn (C. Eickhoff, *et. al.*, 2012) Picture This; (Bennett, *et. al.*, 2009b); FuFinder (O'Neil, Purvis, & Azzopardi, 2011); Book Explorer (Kazai, Milic-Frayling, & Costello, 2009)
11 (U)	Refine information need or query	Refine information need (Step 7) or query (Step 8) based on the evaluation findings in Step 10	See Step 7; however most approaches focus on initial search, not refinements.	Find It If You Can (Ageev, *et. al.*, 2011); Page Hunt (Ma, *et. al.*, 2009); Search War (Edith Law, *et. al.*, 2009); Koru (Milne, *et. al.*, 2008);
Steps 7 to 11 above are repeated until no further refinements are requested				
12 (SYS)	Display final ranked set of relevant documents	Obtain and display the most correct set of relevant documents	No relevant crowdsourcing work found	No relevant games found.

Legend

SD typically completed by the system developer
SYS typically completed by the system
U typically completed by the user

Table 7: Focus of literature on of crowdsourcing and games in IR preprocessing steps

Step No	Step Description	Task	Recent research and work in Crowdsourcing	Recent research and work in Games
3a (SYS)	Perform lexical analysis	Break each document into tokens for analysis	No relevant crowdsourcing work found	No relevant games found.
3b (SYS)	Term resolution	Resolve term acronyms and abbreviations	A number of Entity Recognition tasks, such as CrowdDB (Franklin, Kossmann, Kraska, Ramesh, & Xin, 2011), as well as work by Su *et. al.* (Su, Pavlov, Chow, & Baker, 2007) Robson *et. al.* (Robson, Kandel, Heer, & Pierce, 2011); abbreviation resolution, e.g. Finin (Finin *et. al.*, 2010)	No relevant games found.
3c (SYS)	Tag term parts of speech	Determine and tag the part of speech for each token	No relevant crowdsourcing work found	No relevant games found.
3d (SYS)	Remove stop words	Remove specific tokens from the collection	No relevant crowdsourcing work found	No relevant games found.
3e (SYS)	Stem tokens	Reduce each token to a stemmed form	No relevant crowdsourcing work found	No relevant games found.
3f (SYS)	Analyze tokens	Perform analysis on document tokens as an input into Step 5 (configure system)	No relevant crowdsourcing work found	No relevant games found.
3g (SYS)	Classify documents	Assign documents to one or more classes	News topic classification in blogs (P.G. Ipeirotis, *et. al.*, 2010; R. M. C. McCreadie, *et. al.*, 2010); classification of movies by MPAA rating (Ipeirotis & Paritosh, 2011)	Labeling games that use a set of labels instead of user-defined ones.

Legend

SD typically completed by the system developer
SYS typically completed by the system
U typically completed by the user

CHAPTER 3

TASK ASSESSMENT FRAMEWORK

Our first goal is to develop a framework, consisting of a set of criteria, to determine the suitability of crowdsourcing and games to a given task. Our framework has two aspects; first, we determine if the task can be completed entirely through the use of human computation methods alone; second, we evaluate if human computation be used to augment existing algorithm-based processes.

3.1 Criteria for Using Human Computation Alone

There are two criteria that are mandatory for a task to be implementable using human computation methods alone. These criteria are determined by making a common-sense evaluation of this task as well as a review of relevant literature. Our two mandatory criteria are:

Criterion 1: Can the human computation mechanism (either crowdsourcing or games) handle the scale of the task?

A task could require millions of precise evaluations to be made. This task is performed far more efficiently through algorithms alone. The use of human computation would be detrimental in terms of speed and cost.

To illustrate this criterion using our IR model (Figure 2, page 6), consider one of the preprocessing tasks – stemming tokens (Step 3e). In this Step, we "stem" or reduce inflected or derived terms to a stem, base or root form, usually through the use of a stemming algorithm. To accomplish this Step using the crowd or a game, millions of tokens would need to be evaluated and stemmed in a consistent way. Clearly, algorithms can perform this Step far more efficiently; it would take humans an unacceptable amount of time. Therefore, the stemming Step fails our scalability test. Similarly, all seven of the preprocessing Steps (Steps 3a - 3g) do not scale for crowdsourcing or game design, given the large number of items to

be processed. Similarly, indexing (Step 4) and the ranking and retrieval (Step 9) also fail our scalability test.

Criterion 2: Does the mechanism require specialized or local knowledge to complete?

A task may require extensive local knowledge, such as an understanding of user expectations of the IR system, the existing and expected system capabilities, or other local constraints that neither the crowd nor game players could be made aware of in a reasonable amount of time. In our IR model (Figure 2), domain definition (Step 1) usually requires considerable knowledge of the local system and information needs. Likewise, configuration of the retrieval system application (Step 5) requires specialized knowledge of IR systems, such as the ability to tune and configure parameters and develop a search strategy. Human computation methods do not satisfy this criterion and so these tasks are eliminated from further consideration.

Table 8: Steps from our IR model that meet the each of the first two criteria.

Step Number	Step Description
2	Obtain Document Collection
6	Implement User Interface
7	Identify information need
8	Obtain query terms and operators
10	Evaluate query results against an information need
11	Refine information need or query

Examining our IR model, the steps that satisfy both the scalability criterion and that do not require specialized or local knowledge are given in Table 8. Next, we consider the following criteria:

Criterion 3: Can the outputs provided by the mechanism be efficiently integrated in real-time into the task?

With Criterion 3, we wish to determine if the mechanism outputs can efficiently integrated into the process without adversely affecting the amount of time taken to complete that task.

Examining our IR model, some tasks such as Step 10 (evaluating query results against an information need) and Step 11 (refining an information need or query) need to be completed in real-time or near-real time. This is particularly true in situations where a user is involved or is waiting for results to be returned or refined. If this evaluation takes an unreasonable amount of time, this criterion is not met. Criterion 3 is particularly important for those Steps that involve handling the user query (Steps 7-12).

The following two additional criteria apply only to games and are not applicable to crowdsourcing. Each of these questions is designed to be answered in a "yes"/"no" format.

Criterion 4: Can the mechanism be designed to be entertaining and yet accomplish the objectives of the task?

Criterion 4 determines the potential of making the task engaging or entertaining. It evaluates whether the concept of *flow*, as described by Csikszentmihalyi (Csikszentmihalyi, 1991) can be maintained while still attaining the task's primary objective. To illustrate using our IR model, it would be challenging to implement a user interface (Step 6) as an engaging game.

Criterion 5: Can one design a scoring mechanism to score game players in real time and in a way that is aligned with the task's objective?

Criterion 5 evaluates if it is possible to score and reward performance in "real time" for proper execution in a task. Some tasks, such as the one where the game format obtains relevance feedback on queries (i.e., Step 10 in our IR model), can be scored in real time if the user is involved; other tasks, such as obtaining a document collection (Step 2), require a longer period of time before their outputs can be evaluated and scored, making that Step

unsuitable for a game format. Table 9 summarizes the suitability of applying crowdsourcing and game methods to an IR model.

Table 9: Assessment of the suitability of applying crowdsourcing and games to an IR model

Step	Criteria	3	4	5	Suitability Rating
	Mechanism	**Can Be Integrated**	**Entertaining Yet Meets Objective**	**Scoring Aligns w/ Objective**	
2 – Obtain Document Collection	Crowdsourcing Games	Yes No	N/A No	N/A No	High Low
6 – Implement User Interface	Crowdsourcing Games	Yes Yes	N/A No	N/A No	High Medium
7 – Identify information need	Crowdsourcing Games	Yes Yes	N/A Yes	N/A No	High Medium
8 – Obtain query terms and operators	Crowdsourcing Games	Yes Yes	N/A Yes	N/A No	High Medium
10 – Evaluate query results against an information need	Crowdsourcing Games	Yes Yes	N/A Yes	N/A Yes	High High
11 – Refine information need or query	Crowdsourcing Games	Yes Yes	N/A Yes	N/A No	High Medium

3.2 Criteria for Augmentation with Human Computation

Our initial test examined if each step in our IR model could be accomplished by having either the crowd or the game complete all work associated with the task. An alternative model is where the humans provide a supplemental role, like handling the most difficult components, or providing quality control. In other words, what if we took an approach, where computers and humans each apply their strengths in a hybrid fashion?

Thus we apply "human value-added" criterion on those steps we eliminated for their inability to scale (Criterion 1).

Criterion 6: Can the mechanism provide value by supplementing the current algorithm-based processing for a given task?

Criterion 6 evaluates whether the crowd or game mechanism can be designed to add value to a primarily automated task. This supplementary role is likely to apply in situations where human computation is not sufficient. For example, Step 3b from our IR model is a preprocessing step that involves the resolution of acronyms and abbreviations. If a document collection contains many acronyms or abbreviations, it may be difficult to accomplish resolutions in real time by humans and these are more suitable for algorithm mechanisms than human ones. However, if the acronyms or abbreviations in the collection are rare and/or ambiguous, they may not be easy for an algorithm to resolve. Criterion 6 examines the ability to have the do most of the work, while having human computation do portions of the task the algorithm is unable to accomplish.

Table 10 illustrates those Steps in the IR model that meet criteria 6 and may be appropriate for the augmented model.

Table 10: Steps in our IR model that meet criteria 6.

Step Number	Step Description
3b	Term resolution
3c	Tag term parts of speech
3g	Classify documents

In Tables 11 and 12, we provide an overall assessment of each step of our IR model and preprocessing steps, respectively. Based on the evaluation of these six criteria, we establish our own assessment of none, low, medium or high applicability to both crowdsourcing and games.

The framework we have described identifies several IR tasks that demonstrate potential for either being fully performed through human computation methods or by augmenting algorithms with human computation. In this thesis, we study several of these

tasks and conduct experiments to the applicability of human computation. The IR tasks we will study are the following:

- Term Resolution (Step 3b)
- Obtain query terms and operators for the user's search (Step 8)
- Evaluate query results against information need (Step 10)
- Refine the information need or query (Step 11)

Table 11: Assessment of the application of crowdsourcing and games to the IR model

Step No	Step Description	Task Objective	Assessment of Applying Crowdsourcing	Assessment of Applying Games
1	Define domain	Define the scope of information for retrieval tasks	○○○ Requires local/special knowledge to perform.	○○○ Requires local/special knowledge to perform.
2	Obtain document collection	Determine the documents to be indexed and available for retrieval. Web spidering, web link analysis, etc., should be done at this stage.	●●● High – finding new data sources to apply to existing domains is something the crowd could readily assist with.	●○○ Low – Making this a fun game, is a challenge; making it possible to score in real time would be difficult
3	Preprocess documents	Provide data enrichment	See Table 12 for specific details on each of the preprocessing Steps	
4	Index documents	Create an index for all documents	○○○ Does not scale, little ability to add human value	○○○ Does not scale, little ability to add human value
5	Configure retrieval system	Determine the best search strategies and parameters for an anticipated set of tasks	○○○ Requires local/special knowledge to perform.	○○○ Requires local/special knowledge to perform.
6	Implement user interface	Have an interface that users can use to meet their needs	●●● High – having the crowd help design user interfaces is promising	●●○ Medium – although there is some benefit to using games, the challenge of creating a user interface might lack the excitement needed
7	Identify information need	Identify the user's information need	●●● High – crowdsourcing is particularly useful to help define complex information needs or to assist novice users	●●○ Medium – the challenge to the game format is to allow the game to be scored in real time.
8	Obtain query terms and operators for the user's search	Obtain the initial query terms and any operators (e.g., Boolean) and apply them.	●●● High crowdsourcing is particularly useful to obtain query terms and operators	●●○ Medium – the challenge to the game format is to allow the game to be scored in real time.

Table 11 Continued

Step No	Step Description	Task Objective	Assessment of Applying Crowdsourcing	Assessment of Applying Games
9	Retrieve and rank user query results	Provide a ranked set of relevant documents based on the submitted query	○○○ Does not scale, little ability to add human value	○○○ Does not scale, little ability to add human value
10	Evaluate query results against information need	Assess the ranked results (Step 9) against the information need (Step 7)	●●● High – getting assistance from the crowd to compare results and information need is a crowdsourcable task	●●● High – performing relevance assessments can be done in real time and made interesting.
11	Refine information need / query	Refine information need (Step 7) or query (Step 8) based on the evaluation findings in Step 10	●●● High – having the crowd evaluate or aid in refinement of the information need (Step 8) based on a set of retrieval results (Step 11) is appropriate as a crowd task	●●○ Medium – identifying the information need can be made into an interesting game. The challenge is to make it possible to score it in real time.

Table 12: Assessment of the application of crowdsourcing and games to the preprocessing steps of the IR model

Step No	Step Description	Task Objective	Assessment of Applying Crowdsourcing	Assessment of Applying Games
3a	Perform lexical analysis	Break each document into tokens for analysis	●○○ Low – The error rate on this Step is low, except in specialized domains (such as chemical terms). There is limited human value added.	●○○ Low – The error rate is generally low, though it would be easy to make this into a game. There is limited human value added.
3b	Term resolution	Resolve term acronyms and abbreviations	●●● High – The human value-added component is high, given a substantial algorithm error rate.	●●● High – This could be made into a game that could be fun and evaluated against a lexicon in real time.
3c	Tag term parts of speech	Determine and tag the part of speech for each token	●●○ Medium – This could be evaluated by the crowd. The low algorithm error rate keeps this from being rated high.	●●○ Medium – This is also a task that could be evaluated as a game.
3d	Remove stop words	Remove specific tokens from the collection	●○○ Low – The low error rate limits the human value added. Creating a stop list may be slightly more valuable	●○○ Low – It would be a challenge to turn this into a game, since it involves examining suitable terms in a very large document collection and thus is not practical

Table 12 Continued

Step No	Step Description	Task Objective	Assessment of Applying Crowdsourcing	Assessment of Applying Games
3e	Stem tokens	Reduce each token to a stemmed form	●○○ Low – Stemming follows rigorous rules, and the human value added is low	●○○ Low – It would be difficult to turn a stemming task into a fun game
3f	Analyze tokens	Perform analysis on document tokens as an input into Step 5 (configure system)	●○○ Low – This token analysis usually involves examining aggregate information, which humans provide limited value, particularly since it requires specialized knowledge	●○○ Low – Since it requires specialized knowledge, it would be difficult to make into a game, and difficult to score in "real time"
3g	Classify documents	Assign documents to one or more classes	●●● High – Humans can provide a training set, or perform quality assurance on algorithm-classified documents.	●●● High – Classification is a task that is easy to turn into a game, and thus the human value added is high

Legend for Tables 11 and 12:

●●● strong applicability

●●○ moderate applicability

●○○ weak applicability

CHAPTER 4

RESEARCH QUESTIONS AND EXPERIMENTS

4.1 Research Questions

We evaluate the following three research questions in this thesis.

Research Question 1 – Comparing human computation modes of interaction (game vs. non-game interfaces)

How do game and non-game modes of interaction compare, in terms of performance (quality or cost), in different IR-based tasks?

Research Question 2 – Comparing types of human computation participant types (students vs. crowdworkers)

How do somewhat identifiable and homogenous participants (students) and largely anonymous individuals (crowdworkers) compare, in terms of performance (quality or cost), in different IR-based tasks? We also use algorithms as a baseline.

Research Question 3 – Comparing data collections (general-purpose vs. specialized)

How does the performance of different modes of interaction and participant types vary between general purpose datasets and specialized datasets?

4.2 Experiments

In Chapter 2, we described how these three factors (modes of interaction, participant type, and types of collections) have been examined in the literature, but these were done independently and did not evaluate the interaction effects between them. Likewise, previous studies did not examine the effects of an augmented approach to participant types in an IR context. Using the information we obtain through our experiments, we can gain important insight into when it is most appropriate to use a particular approach.

Our objective is to examine how well human computation can accomplish or assist with tasks from our IR model. From the IR model provided in Figure 2, we select several

strong candidates for further examination: a preprocessing Step involving acronym identification and resolution (Step 3b), evaluating query results against an information need (Step 10), and initial query evaluation (Step 8) and query refinement with feedback (Step 11). These three represent experiments in Chapters 5, 6, and 7, respectively. We present experiments designed to address our research questions in the context of these tasks.

4.3 Metrics

We use the following performance measures in our experiments: precision, recall, and LAM. The matrix in Figure 5 helps define these performance measures.

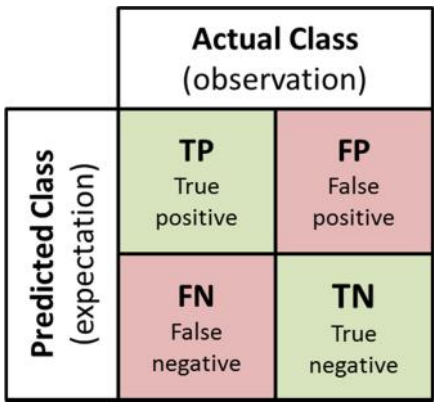

Figure 5: Matrix illustrating the possible classifications of results in a task

Precision, which measures *exactness*, or the ability to retrieve those items that are most important relevant to an information need, is defined as:

$$\frac{TP}{TP + FP}$$

.

If the items we seek are ranked, precision can be evaluated at different depths (e.g., p@10, which measures precision for the top 10 retrieved items).

If a ranked list of items is returned, it is desirable to also consider the order in which the returned items are presented. A precision and recall score can be calculated at every position in the ranked sequence of items. Using this set of calculations, one can plot a precision-recall curve, plotting precision as a function of recall, *r*. *Average precision* (AveP) computes the average value of precision, *p(r)* over the interval from r = 0 to r = 1:

$$AveP = \int_0^1 p(r)dr$$

This is the area under the precision-recall curve. In practice, however, this integral is replaced with a finite sum over every position in the ranked sequence of items:

$$AveP = \sum_{k=1}^{n} p(k)\Delta r(k)$$

where *k* is the rank in the sequence of retrieved items, *n* is the number of retrieved items, *p(k)* is the precision at cut-off *k* in the list, and $\Delta r(k)$ is the change in recall from items *k-1* to *k* (Zhu, 2004). This finite sum is equivalent to:

$$AveP = \frac{\sum_{k=1}^{n} p(k) \times rel(k)}{number\ of\ relevant\ items}$$

where (Meyer, *et. al.* 2012)

$$rel(k) = \begin{cases} 1;\ if\ item\ at\ rank\ k\ is\ relevant \\ 0;\ otherwise \end{cases}$$

Recall, which measures completeness, or ability of the search to find all of the relevant items in the collection, is defined as:

$$\frac{TP}{TP + FN}$$

The *F-score* takes into account both recall and precision. It measures the harmonic mean of precision and recall and is given as follows:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

Compared to the arithmetic mean, both recall and precision need to be high for the harmonic mean to be high.

The *logistic average misclassification* rate (LAM), introduced for the TREC 2005 Spam Filtering track (Cormack & Lynam, 2005), was used as the primary evaluation metric for TREC 12 Crowd TRAT task (Smucker, Kazai, & Lease). LAM is defined as

$$LAM = logit^{-1}\left(\frac{logit\,(fnr) + logit\,(fpr)}{2}\right)$$

where *fnr* is the smoothed false negative rate and the *fpr* is the smoothed false positive rate.

$$fpr = \frac{|FP| + 0.5}{|FP| + |TN| + 1}$$

$$fnr = \frac{|FN| + 0.5}{|FN| + |TP| + 1}$$

The logit function (and its inverse) are defined as:

$$logit\,(p) = \log\frac{p}{1-p}$$

$$logit^{-1} = \frac{e^x}{1-e^x}$$

Thus, lower values of LAM are desirable.

In addition to these metrics, we also examine the *cost* to complete some of our tasks. Our web-based interfaces allow us to examine the time taken in each stage of many of our tasks.

CHAPTER 5

ACRONYM IDENTIFICATION AND RESOLUTION STUDY

Acronyms are used in many document collections to abbreviate and stress important concepts. The identification of acronyms and discovery of their associated definitions are essential aspects to tasks such as natural language processing of texts as well as knowledge-based tasks such as information retrieval, named entity resolution, ontology mapping and question-answering. The identification and resolution of acronyms are not trivial tasks – in many domains, acronyms evolve rapidly. Existing terminological resources and scientific databases cannot keep up-to-date with the growth of these neologisms. Attempts to manually compose large-scale lexicons of acronym-definition pairs suffer from these same challenges, including that such lexicons become obsolete quickly, and there are difficulties in the resolution of ambiguous terms, and variant forms of the same acronym. For example, Acronym Finder, the world's largest dictionary of acronyms, includes more than 1 million human-edited definitions and is growing at an average of 375 per day; the total effort required to compile this set is estimated to be more than 9,000 hours, a task performed during the last 14 years (Molloy, 2010).

Attending to these shortcomings with many current approaches, this experiment evaluates several different non-lexicon approaches (e.g., (Larkey, Ogilvie, Price, & Tamilio, 2000; Park & Byrd, 2001; Sanchez & Isern, 2011) to identify and resolve short-form acronyms and their definitions – using a variety of different approaches. This study was conducted over 11 days in late May and early June, 2013.

5.1 Review of Acronym Identification and Resolution Literature

A significant percentage of the acronym detection and resolution literature relate to evaluating acronyms in medical text. We evaluate these methods on two text collections with different characteristics: a collection of News articles and a collection of patents. Human-

based methods possess advantages that algorithm methods do not; therefore our focus is also on examining where human-based methods provide value, and where they do not, in acronym identification and resolution. The goal of acronym identification and resolution is to extract pairs of short forms (acronyms) and long forms (their expanded forms or definitions) occurring in text. Much of the work with acronyms is either limited to a specific domain (e.g., biomedical text or government documents) or requires the algorithm to be trained on the corpus before use.

In 2003, the TREC Genomics track (Hersh & Bhupatiraju, 2003) began a task that invited acronym identification and definition extraction in biomedical text. This TREC-motivated research encouraged the development of a number of algorithms that performed well against biomedical text it was designed to handle. However, few methods used to examine biomedical text have also demonstrated their ability to work effectively on text in other domains. There have been some attempts to use the broader web to extract definitions of terms, such as that by Sanchez & Isern (Sanchez & Isern, 2011) Their method is language independent, but is reliant on a large corpus for acronym resolution and may not scale well for documents with complex or rarely-occurring acronyms.

One challenge in acronym identification and resolution is there are no rules or precise patterns for the creation of acronyms. Moreover, acronyms are ambiguous – the same acronym may refer to two or more different concepts (e.g., IEM abbreviates both “immune-electron microscopy” and “interstitial electron model”) and have variant forms (e.g. an example provided by Buttcher et.al. in (Buttcher, Clarke, & Cormack, 2004), the Medline corpus refers to the “NF-kappaB” protein in 6 different ways: “NF-kappa B” (33902), “NF-kappaB” (28551), “NFkappaB” (3211), “NF-kB (688), “NFkB” (259), and “NFkappa B” (45)). Ambiguity and variation present several challenges in text mining approaches, since acronyms have not only have to be recognized, but their variants have to be linked to the same canonical form and be disambiguated, adding to the complexity of acronym recognition through the use of algorithms.

Yeates, Bainbridge and Whitten (Yeates, Bainbridge, & Witten, 2000) built an acronym detection scheme using 150 reports from the Computer Science Technical Reports collection of the New Zealand Digital Library. They identify acronyms using a rigidly-defined set of rules achieving reasonable precision using a "compression technique". In a preliminary investigation using their parameters, we found that their prescribed rules would not achieve the same results with our datasets. Xu and Huang approach this problem using SVMs (Xu & Huang, 2007). Likewise, Feng *et. al.*, use an unsupervised and graphing algorithms (Feng, Xiong, Yao, Zheng, & Liu, 2009). More recently, Yarygina and Vassilieva (2012) address acronym detection using a relevance feedback technique. Like other methods, they first obtain a list of acronyms with high recall but low precision rates. Definitions are constructed for every short-form acronym candidate from its surrounding text. Next, a classifier is used to select genuine long-form/short-form pairs. Last, they apply relevance feedback and claim reasonable precision without losing recall. The disadvantage of these methods is they all require training or they require knowledge of the underlying data collection.

Schwartz and Hearst (2002) implemented an algorithm for identifying acronyms. Unlike the methods described earlier, their method does not need prior training; instead they use parenthetical expressions as a marker of a short form. An emphasis is on complicated acronym-definition patterns for cases in which only a few letters match. Despite the core algorithm being admittedly simple, the authors report 99% precision and 84% recall on MEDLINE abstracts. As a result of its simplicity and ease of implementation, this algorithm appears appropriate for generalization to collections outside of biomedicine. Dannélls (Dannélls, 2006) provided a modified version of the Schwartz and Hearst algorithm, with the advantage of recognizing acronym-definition pairs not indicated by parentheses. They were able to achieve good precision (above 90%) and recall (above 96%) against four Swedish medical text collections. (Dannélls; Kokkinakis & Dannélls, 2006; Yarygina & Vassilieva, 2012).

Human based approaches can add considerable value to acronym identification and resolution. Humans are able to adapt to the non-standardized rules commonly found in acronym identification and make use of outside knowledge and apply this to acronym resolution. Despite this, no work has been found in the literature that examines crowdsourcing's ability to detect and resolve acronyms, although some other studies have examined the reliability of the crowd for named entity resolution (NER) with Twitter data (Finin, *et. al.*, 2010), annotating images (Stefanie Nowak & Ruger, 2010), and evaluating document relevance with a set of prescribed TREC topics (Omar Alonso & Mizzaro, 2012). In this experiment, we explore the value the crowd provides in acronym identification and resolution.

5.2 Contributions

This study examines the acronym resolution preprocessing Step, which is Step 3(b) of our IR model (Figure 2). As indicated in Chapter 3.3, this Step does not meet the initial criteria for using human computation methods alone. However our framework does indicate the potential for a human value-added contribution.

There are two situations we explore: to identify acronyms (locating acronyms in text) and acronym definition (finding the long-form definition of a given acronym) in text collections. These represent Phase 1 and Phase 2 of our experiment, respectively.

This experiment offers the following. First, we take a string-matching algorithm that has demonstrated good results in one domain and test its effectiveness on two new domains. Second, we examine the effectiveness of two human computation approaches – one approach uses a game-based mode of interaction and another uses crowd participants. Last, we study the value of using two different types of human participants, a largely anonymous group of people and a more homogenous and not so anonymous group of individuals; namely students in a campus.

5.3 Hypothesis

We examine the following six hypotheses related to acronym identification and acronym resolution. We examine three factors: interface type (game and non-game), participant type (crowd and student) and collection type (News and Patent). We measure performance by examining precision, recall and F-score.

5.3.1 Acronym Identification Precision

$H_{0\text{-}1P}$: there are no main or interaction differences in mean precision on acronym identification due to the type of interface, type of participant, or type of collection used.

$H_{A\text{-}1P}$: there are main or interaction differences in mean precision on acronym identification due to the type of interface, type of participant, or type of collection used.

5.3.2 Acronym Identification Recall

$H_{0\text{-}1R}$: there are no main or interaction differences in mean recall on acronym identification due to the type of interface, type of participant, or type of collection used.

$H_{A\text{-}1R}$: there are main or interaction differences in mean recall on acronym identification due to the type of interface, type of participant, or type of collection used.

5.3.3 Acronym Identification F-score

$H_{0\text{-}1F}$: there are no main or interaction differences in mean F-score on acronym identification due to the type of interface, type of participant, or type of collection used.

$H_{A\text{-}1F}$: there are main or interaction differences in mean F-score on acronym identification due to the type of interface, type of participant, or type of collection used.

5.3.4. Acronym Resolution Precision

$H_{0\text{-}2P}$: there are no main or interaction differences in mean precision on acronym definition resolution due to the type of interface, type of participant, or type of collection used.

$H_{A\text{-}2P}$: there are main or interaction differences in mean precision on acronym definition resolution due to the type of interface, type of participant, or type of collection used.

5.3.5. Acronym Resolution Recall

$H_{0\text{-}2R}$: there are no main or interaction differences in mean recall on acronym definition resolution due to the type of interface, type of participant, or type of collection used.

$H_{A\text{-}2R}$: there are main or interaction differences in mean recall on acronym definition resolution due to the type of interface, type of participant, or type of collection used.

5.3.6 Acronym Resolution F-score

$H_{0\text{-}2F}$: there are no main or interaction differences in mean F-score on acronym definition resolution due to the type of interface, type of participant, or type of collection used.

$H_{A\text{-}2F}$: there are main or interaction differences in mean F-score on acronym definition resolution due to the type of interface, type of participant, or type of collection used.

5.4 Collections

General Document Collection: News article documents from TREC collection, TREC disks 4 and 5, minus the Congressional Record. The documents in this collection are small News articles from the Financial Times, Federal Register, LA Times, and Foreign Broadcast Information Service, covering 1989 through 1996. Documents were an average of 1055 words in length, with a Flesch-Kinkaid reading level of 12.17 and an Automated Readability Index (ARI) of 12.64. We used acronyms from the headline and text fields only.

Specialized Document Collection: Patent documents obtained from the Matrixware Research Collection (MAREC), which includes European, Japanese, and US patents, from

2001-2008 (Oostdijk, Verberne, & Koster, 2010). We used acronyms from patent's title, abstract, field of the invention and background of the invention sections of the description. In the Patent data collection, due to the longer length of the documents involved, we limit our use to the first 1000 words from each document, rounded to the nearest paragraph. We do this in order to keep the collections comparable. Documents were an average of 1070 words in length, with a Flesch-Kinkaid reading level of 13.03 and an Automated Readability Index (ARI) of 13.31.

All documents in both collections were in English. We take 20 randomly-selected documents (40 documents total) from each of two text collections. Table 13 provides some background on the text characteristics in each collection.

Table 13: Readability indices and text characteristics for the News (n=20) and Patent (n=20) collections

	News		**Patent**	
Readability Indices	**Mean**	**SD**	**Mean**	**SD**
Flesch Kincaid Reading Ease	45.38	10.06	40.01	14.06
Flesch Kincaid Grade Level	12.17	3.71	13.03	4.20
Automated Readability Index	12.64	3.93	13.31	5.76
Text Characteristics	**Mean**	**SD**	**Mean**	**SD**
No. of acronyms	6.20	3.28	6.55	4.49
No. of acronym definitions	5.10	2.38	4.90	2.69
No. of sentences	51.90	18.69	59.95	27.30
No. of words	1041.60	187.02	1059.70	84.17
No. of complex words	184.00	12.20	186.95	14.03
Percent of complex words	17.4%	3.3%	17.5%	3.4%
Average words per sentence	21.08	7.23	21.57	9.70
Average syllables per word	1.66	0.10	1.71	0.10

Since we are comparing performance across collections, we wish to ensure that the collections are as similar as possible. We establish their similarity as follows. Using two-tailed t-tests, we evaluate the following hypotheses for each of 11 characteristics:

$H_{0\text{-Coll}}$ = the means of the two collections are the same

$H_{A\text{-}Coll}$ = the means of the two collections are different

Table 14 provides information on the differences between the two collections. From this analysis, we note that for all 11 characteristics, the p-value > 0.05, therefore, we cannot reject $H_{0\text{-}Coll}$ and we cannot assume that any of the text characteristics examined are different between the two collections.

Table 14: Statistical significance of readability indices and text characteristics for the News and Patent collections

Readability Indices	**df**	**t**	**p**	**Significant @ p<0.05?**
Flesch Kincaid Reading Ease	38	1.3891	0.1729	No
Flesch Kincaid Grade Level	38	0.6863	0.4967	No
Automated Readability Index	38	0.4297	0.6698	No
Text Characteristics	**df**	**t**	**p**	**Significant @ p<0.05?**
No. of acronyms	38	0.2815	0.7799	No
No. of acronym definitions	38	0.2490	0.8047	No
No. of sentences	38	1.0881	0.2834	No
No. of words	38	0.3947	0.6967	No
No. of complex words	38	0.7096	0.4823	No
Percent of complex words	38	0.0094	0.9253	No
Average words per sentence	38	0.1811	0.8572	No
Average syllables per word	38	1.5811	0.1221	No

5.5 Gold standard

To create a gold standard acronym list, we had two human assessors each independently evaluate the 20 News and 20 Patent documents to both identify acronyms and identify their related expansions. Each acronym and its definition, if available, was counted only once per document, and lists for each assessor for each document were adjudicated over any disagreements. Assessors indicated if the acronym's definition was available in the document, whether it was available through external common knowledge, or neither. The gold standard acronym count for each collection is provided in Table 15 and the Cohen's Kappa statistic, which indicates inter rater reliability, for the identification and resolution phases for each collection. This information is provided in Table 16.

Table 15: Gold standard acronym count

Text Collection	Acronym Identification				Acronym Resolution	
	Assessor 1	Assessor 2	Adjudicated Agreement on Identified Acronyms	Adjudicated Acronyms Requiring External Common Knowledge to Identify	Adjudicated Agreement on Resolved Definitions	Adjudicated Acronyms Requiring External Common Knowledge to Resolve
News	136	135	130	15	127	44
Patent	137	138	127	14	125	23

Table 16: Cohen's kappa statistics for each collection

Collection	Identification	Resolution
News	0.771	0.634
Patent	0.725	0.552

Using the interpretation set by Landis and Koch (Landis & Koch, 1977), the inter-annotator agreement for both acronym identification and resolution are substantial. This provides confidence that the acronyms identification and resolution obtained from both collections are accurate.

5.6 Interaction Modes & Baseline

5.6.1 Algorithm Baseline

Our *algorithm baseline* is the language-independent enhancement (Dannélls, 2006) of Schwartz and Hearst's acronym detection algorithm (Schwartz & Hearst, 2002). Dannélls algorithm is a good candidate for our study since it, along with Schwartz and Hearst's algorithm, has been studied fairly extensively in the literature, does not require supervision or training, and is not lexicon-dependent. This algorithm was described in Section 5.1. The

pseudo code is provided in Appendix C and the perl code used to implement this algorithm is provided in Appendix D.

5.6.2 Human Computation Modes of Interaction

We use two modes of interaction – a non-game interface, which solicits input through a PHP-based Web 2.0 interface and game interface, which uses a PHP-based game. For each of these modes of interaction, participants were shown the same 20 News or 20 Patent documents in precisely the same order in each phase of this experiment.

5.6.2.1 Non-game Interface

In the first phase (acronym identification), the non-game interface displays a document and asks the user to write each acronym they find in the document into a text box. In the second phase (acronym resolution), the non-game interface highlights the gold standard acronyms in the text, and asks the user to provide a definition of each acronym. The user is also asked to provide the source of the acronym definition –the document or the user's own knowledge. Screenshots for Phase 1 (acronym identification) and Phase 2 (acronym resolution) of the non-game interface are shown in Figure A-1 and Figure A-3, respectively, in Appendix A.

5.6.2.2 Game Interface

The game interface supports the same two phases presented in the web interface; however, their responses are timed and they receive feedback in the form of a score after each document is evaluated. Also, game participants are awarded badges for high performance and are given the option of adding their name to a leaderboard if they have a high score for that phase. Screenshots for Phase 1 (acronym identification) and 2 (acronym resolution) of the game interface are shown in Figure A-2 and Figure A-4, respectively, in Appendix A. Instructions to participants were the same for the game and non-game versions. The text of these instructions is provided in Appendix B.

5.7 Participants

One hundred and forty-four participants completed both phases of this task. Seventy-two of the participants were current students at a U.S. or Canadian institution. These student participants were required to affirm they were current students. Each was required to have a valid ".edu" email address in order to participate. Seventeen different educational institutions were represented. Students who completed both phases were paid $10.00, a typical compensation for time and effort in study participation. Seventy-two crowd participants were recruited using Amazon Mechanical Turk and were paid $0.60 to complete both phases. Crowdworkers participated from 27 different countries (as determined by IP address).

Table 17: Demographic information obtained from participants through a survey. Percentage indicating each choice is given in parentheses.

Category	**Participant Response**
Region (as determined by IP address)	North America (56.9%), Europe (16.0%), Africa/Middle East (2.1%), South Asia (18.1%), East Asia (4.2%), South America (0.7%), Australia/NZ/Oceana (2.1%)
Age	<18 (15.3%), 19-25 (59.0%), 26-35 (21.5%), 36-45 (3.5%), 46+ (0.7%)
Gender	Male (45.1%), Female (54.9%)
English Ability	Poor (1.4%), Moderate (13.9%), Good (27.8%), Fluent (56.9%)
Education	No baccalaureate (68.1%), Completed baccalaureate (31.9%)
Current Student?	Yes-Full-time (59.7%) , Yes-Part time (31.3%), No (9.0%)
Chemistry course in last 5 years?	Yes (50.7%), No (49.3%)

Each participant was randomly assigned to a mode of interaction (either the game interface or the non-game interface) and to a collection (either the News collection or the Patent collection). We had 18 participants for each combination of collection, interface, and participant type.

In addition to recording their IP address for geo-location purposes, we also required participants to provide some information. Table 17 provides the responses to each survey categories asked of participants and the percentages responding to each choice.

5.8 Procedure

5.8.1 Acronym Identification (Phase 1)

Algorithm: We ran the algorithm on the 20 documents from each collection, which output the unique acronyms identified and their resolutions. Resolutions not found are marked 'unknown'. For example, common acronyms such as "PM" to represent afternoon, are rarely expanded in documents. Thus we can track errors in terms of not identifying strings that are acronyms. We can also identify errors in resolution.

Non-game: Using our web interface, we had participants read each document and identify the acronyms. For each collection, all participants evaluated the same 20 documents in the same order. Participants could cut and paste text from the document into a textbox. Users were not provided any feedback on their performance during the task.

Game: The game was designed to provide players with a more entertaining and challenging method of identifying acronyms. Using the same documents players had to identify the acronyms within a specified time limit (3 minutes per document), and were given real-time scores at the end of each document and a bonus for quick resolution. Participants could cut and paste text from the document into a textbox. Game participants were given the same instructions as non-game participants. The game interface had upbeat music and a countdown timer to indicate the passage of time. A leaderboard was shown to all game participants at the beginning and at the end of the game for top scorers to enter their names.

5.8.2 Acronym Resolution (Phase 2)

Algorithm: Acronym resolution was done during the same Step as acronym identification. The analysis was limited to the gold standard acronyms identified (124 for the News collection and 131 for the Patent collection).

For human participants, definitions that were misspelled were not considered a valid match with few exceptions (e.g., “defence” in place of “defense”); case and punctuation were not considered.

Non-game: We provided workers the same 20 documents in the same order but with the gold standard acronyms highlighted in the document through our web interface. Participants were asked to resolve each of them. In addition, we asked each worker to mark whether they used information solely from the document, or they used external common knowledge to resolve the acronym definitions. Participants could cut and paste text from the document into a textbox. Users were not provided any feedback on their performance during the task.

Game: As with the acronym identification Step, the game was designed to provide players with a more entertaining and challenging method of resolving acronyms. Using the same highlighted documents players had to resolve the terms within a specified time limit (3 minutes per document), and were given real-time scores at the end of each document. Game participants were given the same instructions as non-game participants. Participants could cut and paste text from the document into a textbox. A bonus could be earned by resolving acronyms quickly and was a percentage of the score earned for that round. The game interface had upbeat music and a countdown timer to indicate the passage of time. A leaderboard was provided for at the beginning and at the end of the game for top scorers to enter their names.

5.9 Results

5.9.1 Acronym Identification (Phase 1)

Table 18 provides the means and standard deviations for our three dependent variables in each experimental condition.

Table 18: Acronym identification means and standard deviations for dependent variables by interface type, collection type and participant type (n = 144)

	Acronym Identification						
	Precision		Recall		F-score		
Condition	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**N**
Interface type							
Game	0.990	0.020	0.603	0.158	0.735	0.160	72
Non-game	0.969	0.048	0.643	0.117	0.767	0.094	72
Collection type							
News	0.975	0.039	0.606	0.129	0.738	0.120	72
Patent	0.985	0.037	0.640	0.149	0.764	0.142	72
Participant type							
Crowd	0.979	0.039	0.615	0.145	0.745	0.131	72
Student	0.981	0.037	0.631	0.135	0.757	0.133	72
Interface × Collection							
Game, News	0.987	0.015	0.617	0.150	0.746	0.149	36
Game, Patent	0.994	0.023	0.590	0.167	0.724	0.172	36
Non-game, News	0.963	0.051	0.595	0.104	0.731	0.084	36
Non-game, Patent	0.976	0.045	0.691	0.110	0.803	0.090	36
Interface × Participant							
Game, Crowd	0.989	0.024	0.617	0.157	0.747	0.152	36
Game, Student	0.992	0.014	0.590	0.160	0.723	0.169	36
Non-game, Crowd	0.969	0.048	0.613	0.134	0.743	0.109	36
Non-game, Student	0.970	0.049	0.673	0.088	0.792	0.069	36
Collection × Participant							
News, Student	0.982	0.027	0.634	0.144	0.758	0.143	36
News, Crowd	0.968	0.048	0.578	0.106	0.719	0.090	36
Patent, Student	0.980	0.045	0.628	0.127	0.756	0.124	36
Patent, Crowd	0.990	0.025	0.652	0.169	0.771	0.160	36

Table 18 Continued

	Acronym Identification						
	Precision		Recall		F-score		
Condition	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**N**
Interface × Collection × Participant							
Game, News, Student	0.986	0.017	0.596	0.176	0.723	0.184	18
Game, News, Crowd	0.987	0.014	0.638	0.120	0.768	0.103	18
Game, Patent, Student	0.997	0.007	0.584	0.147	0.723	0.158	18
Game, Patent, Crowd	0.991	0.032	0.596	0.189	0.725	0.190	18
Non-game, News, Student	0.977	0.035	0.673	0.092	0.794	0.073	18
Non-game, News, Crowd	0.949	0.061	0.518	0.031	0.669	0.031	18
Non-game, Patent, Student	0.963	0.060	0.673	0.086	0.789	0.066	18
Non-game, Patent, Crowd	0.988	0.015	0.709	0.129	0.818	0.109	18

5.9.1.1 Acronym Identification Precision

Table 19: ANOVA results for precision of acronym identification

	Precision				
Effect	**SS**	**df**	**Mean Square**	**F**	**p-value**
Interface	0.016	1	0.016	12.195	0.001
Participant	0.000	1	0.000	0.102	0.750
Collection	0.004	1	0.004	2.788	0.097
Interface × Collection	0.000	1	0.000	0.257	0.613
Interface × Participant	0.000	1	0.000	0.012	0.913
Collection × Participant	0.005	1	0.005	3.610	0.060
Interface × Collection × Participant	0.008	1	0.008	6.150	0.014
Error	0.176	136	0.001		

Table 19 provides the ANOVA results for precision for the acronym identification task. The three-way ANOVA results for precision of acronym identification were statistically significant ($p < 0.05$) for interface, with those using the game interface outperforming those using the non-game interface, $F(1,136)=12.195$, $p=0.001$, for the interface x participant type

x collection interaction, F(1,136)=6.15, p=0.014. This three-way interaction effect is graphically depicted in Figure 6. Simple effect follow-up tests for the three-way interaction revealed that precision was higher for the game interface than for the non-game interface for crowd participants evaluating the News collection (M_G=0.987, M_{NG}= 0.949; M_{Diff} = 0.038; t(136)=3.623; p<0.001; d=1.21) and for students evaluating the Patent collection (M_G= 0.997; M_{NG}=0.963; M_{Diff} =0.033; t(136)= 3.171, p=0.002; d=1.06).

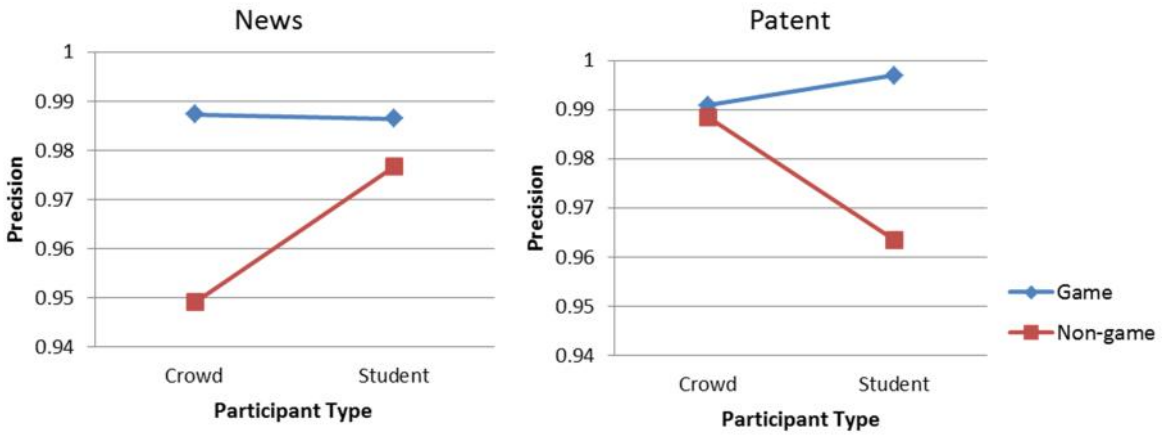

Figure 6: Interaction effects for precision in acronym identification for interface (lines) and participant type (x-axis) for the News collection (left) and Patent collection (right)

5.9.1.2 Acronym Identification Recall

Table 20 provides the ANOVA results for recall for the acronym identification task. The three-way ANOVA results for recall of acronym identification were statistically significant ($p < 0.05$) for the interaction between type of interface and collection, F(1,136)=7.961, p=0.005, between interface and participant, F(1,136)=4.017, p=0.047 and the three-way interaction interface x participant type x collection interaction, F(1,136)=6.537, p=0.012. This three-way interaction effect is graphically depicted in Figure 7.

Table 20: ANOVA results for recall of acronym identification

Effect	Recall				
	SS	df	Mean Square	F	p-value
Interface	0.057	1	0.057	3.345	0.070
Participant	0.010	1	0.010	0.560	0.456
Collection	0.042	1	0.042	2.485	0.117
Interface × Collection	0.135	1	0.135	7.961	0.005
Interface × Participant	0.068	1	0.068	4.017	0.047
Collection × Participant	0.058	1	0.058	3.404	0.067
Interface × Collection × Participant	0.111	1	0.111	6.537	0.012
Error	2.312	136	0.017		

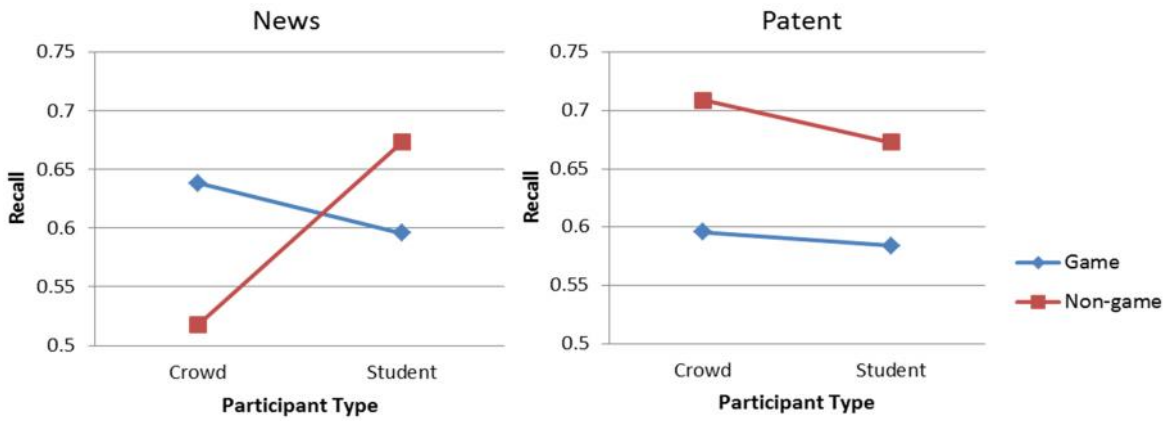

Figure 7: Interaction effects for recall in acronym identification for interface (lines) and participant type (x-axis) for the News collection (left) and Patent collection (right)

Simple effect follow-up tests for the three-way interaction revealed that recall was higher for the game interface than for the non-game interface for crowd participants evaluating the News collection (M_G=0.638, M_{NG}= 0.518; M_{Diff} = 0.121; t(136)= 2.777; p=0.006; d=0.926) .

For the Patent collection, the opposite was found to be true: crowd participants evaluating the Patent collection using the non-game interface obtained higher recall scores

than crowd participants using the game interface (M_{NG}= 0709; M_G=0.596; M_{Diff} =0.113, t(136)= 2.602; p=0.010; d=0.867). Likewise, students using the non-game interface also performed better than students using the game interface (M_{NG}= 0.673; M_G=0.584; M_{Diff} =0.084; t(136)= 2.049, p=0.042; d=0.683).

5.9.1.3 Acronym Identification F-score

Table 21: ANOVA results for F-score of acronym identification

Effect	F-score				
	SS	**df**	**Mean Square**	**F**	**p-value**
Interface	0.038	1	0.038	2.400	0.124
Participant	0.005	1	0.005	0.336	0.563
Collection	0.023	1	0.023	1.430	0.234
Interface × Collection	0.080	1	0.080	4.972	0.027
Interface × Participant	0.047	1	0.047	2.936	0.089
Collection × Participant	0.027	1	0.027	1.706	0.194
Interface × Collection × Participant	0.087	1	0.087	5.430	0.021
Error	2.175	136	0.016		

Table 21 provides the ANOVA results for F-score for the acronym identification task. The three-way ANOVA results for recall of acronym identification were statistically significant ($p < 0.05$) for the interaction between type of interface and collection, F(1,136)=4.972, p=0.027, and the three-way interaction interface x participant x collection interaction F(1,136)=5.43, p=0.021. This interaction effect is graphically depicted in Figure 8.

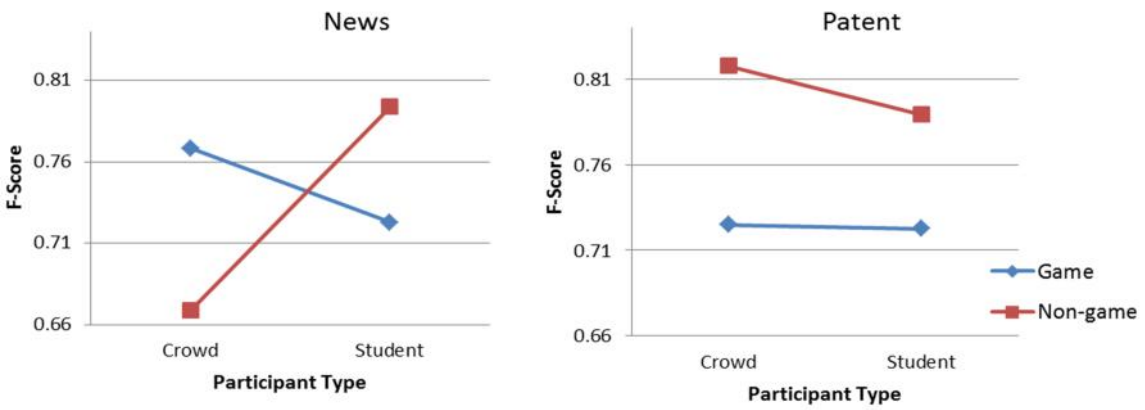

Figure 8: Interaction effects for F-score in acronym identification for interface (lines) and participant type (x-axis) for the News collection (left) and Patent collection (right)

Simple effect follow-up tests for the three-way interaction revealed that F-score was higher for the game interface than for the non-game interface for crowd participants evaluating the News collection (M_G=0.768, M_{NG}=0.669; M_{Diff} = 0.100; t(136)= 2.361; p=0.020; d=0.787). The opposite was true for crowd participants evaluating the Patent collection: those using the non-game interface outperformed those using the game interface (M_{NG}= 0.818; M_G=0.725; M_{Diff} =0.093; t(136)= 2.197; p=0.030; d =0.732).

5.9.2. Resolution (Phase 2)

Table 22 provides the means and standard deviations across our three dependent variables.

Table 22: Acronym resolution means and standard deviations for dependent variables by interface, collection and participant type (n = 144)

	Acronym Resolution						
	Precision		Recall		F-score		
Condition	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**N**
Interface							
Game	0.946	0.077	0.430	0101	0.587	0.110	72
Non-game	0.951	0.057	0.433	0.099	0.590	0.101	72
Collection type							
News	0968	0.027	0.498	0.069	0.655	0.064	72
Patent	0.929	0.088	0.366	0.080	0.522	0.097	72
Participant type							
Crowd	0.946	0.084	0.417	0.083	0.576	0.096	72
Student	0.951	0.047	0.446	0.112	0.601	0.114	72
Interface × Collection							
Game, News	0.966	0.028	0.508	0.044	0.665	0.041	36
Game, Patent	0.926	0.102	0.353	0.080	0.508	0.102	36
Non-game, News	0.970	0.027	0.378	0.087	0.645	0.080	36
Non-game, Patent	0.933	0.072	0.487	0.078	0.536	0.091	36
Interface × Participant							
Game, Crowd	0.943	0.097	0.430	0.095	0.587	0.108	36
Game, Student	0.949	0.052	0.430	0.108	0.586	0.114	36
Non-game, Crowd	0.948	0.069	0.403	0.068	0.564	0.081	36
Non-game, Student	0.954	0.042	0.462	0.116	0.616	0.113	36
Collection × Participant							
News, Student	0.968	0.027	0.542	0.051	0.694	0.053	36
News, Crowd	0.968	0.028	0.453	0.056	0.615	0.048	36
Patent, Student	0.934	0.056	0.351	0.065	0.508	0.078	36
Patent, Crowd	0.924	0.112	0.351	0.090	0.536	0.112	36
Interface × Collection × Participant							
Game, News, Student	0.964	0.031	0.518	0.048	0.674	0.048	18
Game, News, Crowd	0.969	0.025	0.497	0.038	0.656	0.033	18
Game, Patent, Student	0.933	0.063	0.342	0.074	0.498	0.090	18
Game, Patent, Crowd	0.918	0.132	0.363	0.088	0.518	0.113	18
Non-game, News, Student	0.972	0.023	0.566	0.042	0.715	0.039	18
Non-game, News, Crowd	0.967	0.031	0.409	0.029	0.574	0.034	18
Non-game, Patent, Student	0.936	0.049	0.359	0.057	0.517	0.064	18
Non-game, Patent, Crowd	0.929	0.090	0.398	0.092	0.554	0.111	18

5.9.2.1 Acronym Resolution Precision

Table 23 provides the ANOVA results for precision for the acronym resolution task. The three-way ANOVA results for precision of acronym resolution were statistically significant ($p < 0.05$) for the main effect of collection, with those assigned to the News collection users outperforming those assigned the Patent collection, $F(1,136)=12.511$, $p=0.001$. There were no statistically significant interaction effects found for precision in acronym resolution.

Table 23: ANOVA results for precision of acronym resolution

	Precision				
Effect	**SS**	**df**	**Mean Square**	**F**	**p-value**
Interface	0.001	1	0.001	0.219	0.640
Participant	0.001	1	0.001	0.251	0.617
Collection	0.055	1	0.055	12.511	0.001
Interface × Collection	0.000	1	0.000	0.029	0.864
Interface × Participant	0.000	1	0.000	0.001	0.973
Collection × Participant	0.001	1	0.001	0.206	0.651
Interface × Collection × Participant	0.001	1	0.001	0.169	0.682
Error	0.596	136	0.004		

5.9.2.2 Acronym Resolution Recall

Table 24 provides the ANOVA results for recall for the acronym resolution task. The three-way ANOVA results for recall of acronym resolution were statistically significant ($p < 0.05$) for the main effects of participant, with students outperforming the crowd, $F(1,136)=8.122$, $p=0.005$ and collection, with those assigned to the News collection users outperforming those assigned the Patent collection, $F(1,136)=160.681$, $p<0.001$. All two-way interactions between factors were significant. There was also a significant three-way interaction interface x participant type x collection interaction, $F(1,136)=13.730$, $p<0.001$. This three-way interaction effect is graphically depicted in Figure 9.

Table 24: ANOVA results for recall of acronym resolution

Effect	Recall				
	SS	df	Mean Square	F	p-value
Interface	0.000	1	0.000	0.068	0.795
Participant	0.032	1	0.032	8.122	0.005
Collection	0.628	1	0.628	160.681	0.000
Interface × Collection	0.019	1	0.019	4.915	0.028
Interface × Participant	0.031	1	0.031	7.933	0.006
Collection × Participant	0.127	1	0.127	32.551	0.000
Interface × Collection × Participant	0.054	1	0.054	13.730	0.000
Error	0.531	136	0.004		

Simple effect follow-up tests for the three-way interaction revealed that recall was higher for the game interface than for the non-game interface for crowd participants evaluating the News collection (M_G=0.497, M_{NG}= 0.409; M_{Diff} = 0.088; t(136)= 4.189; p<0.001; d=1.397); however, students demonstrated higher recall for the non-game interface than the game interface in the News collection (M_{NG}=0.566, M_G= 0.518; M_{Diff} = 0.0476; t(136)= 2.256; p=0.006; d=0.752).

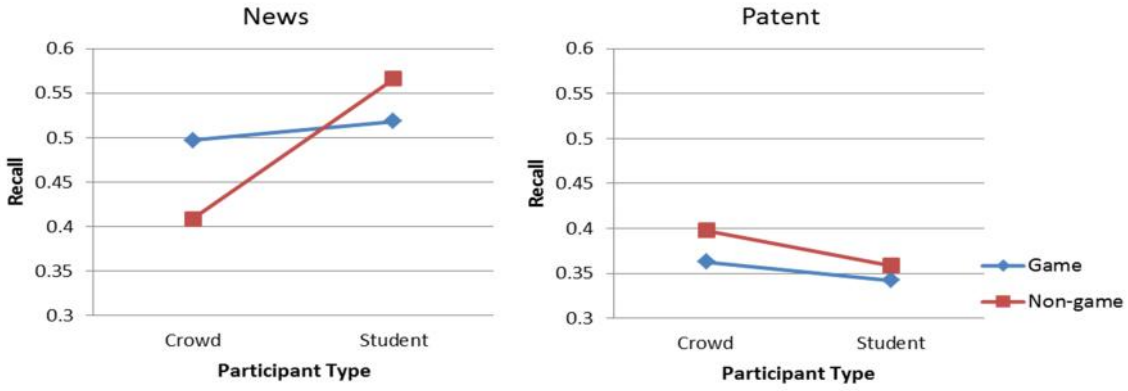

Figure 9: Interaction effects for recall in acronym resolution for interface (lines) and participant type (x-axis) for the News collection (left) and Patent collection (right)

5.9.2.3 Acronym Resolution F-score

Table 25 provides the ANOVA results for F-score for the acronym resolution task. The three-way ANOVA results for recall of acronym resolution were statistically significant ($p < 0.05$) for the main effect of collection with those assigned to the News collection users outperforming those assigned the Patent collection, $F(1,136)=116.849$, $p<0.001$ and for participant, with students outperforming the crowd, $F(1,136)=4.253$, $p=0.041$, for the two-way interaction between type of collection and participant, $F(1,136)=19.000$, $p<0.001$, and interface and participant $F(1,136)=4.679$, $p=0.032$ and for the three-way interaction interface x participant type x collection interaction $F(1,136)=8.082$, $p=0.005$. This three-way interaction effect is graphically depicted in Figure 10.

Table 25: ANOVA results for F-score of acronym resolution

	F-score				
Effect	**SS**	**df**	**Mean Square**	**F**	**p-value**
Interface	0.001	1	0.001	0.092	0.762
Participant	0.023	1	0.023	4.253	0.041
Collection	0.636	1	0.636	116.849	0.000
Interface × Collection	0.103	1	0.103	3.797	0.053
Interface × Participant	0.025	1	0.025	4.679	0.032
Collection × Participant	0.103	1	0.103	19.000	0.000
Interface × Collection × Participant	0.044	1	0.044	8.082	0.005
Error	0.740	136	0.005		

Simple effect follow-up tests for the three-way interaction revealed that the F-score was higher for the game interface than for the non-game interface for crowd participants evaluating the News collection ($M_G=0.656$, $M_{NG}=0.574$; $M_{Diff} = 0.071$; $t(136)= 3.469$; $p=0.001$; $d=1.156$).

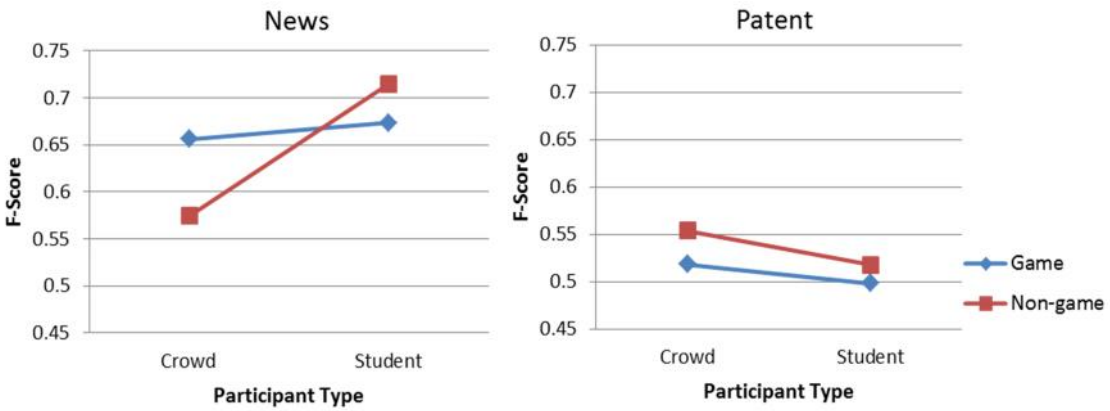

Figure 10: Interaction effects for F-score in acronym resolution for interface (lines) and participant type (x-axis) for the News collection (left) and Patent collection (right)

5.9.3 Summary of Findings and Evaluation of Hypotheses

Table 26 provides a summary of our findings for the Acronym Identification and Acronym Resolution tasks. From Table 26, we can assess our hypotheses. This analysis is provided in Table 27.

Table 26: Summary of findings for the acronym identification and acronym resolution tasks

Effect Type	Acronym Identification			Acronym Resolution		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
Main	Interface: Game > Non-game	None	None	Collection: News > Patent	Collection; News > Patent; Participant: Student > Crowd	Collection; News > Patent; Participant: Student > Crowd
Interaction	Crowd & News: Game > Non-game; Student & Patent: Game > Non-game	Crowd & News: Game > Non-game; Crowd & Patent: Non-game > Game; Student & News: Non-game > Game	Crowd & News: Game > Non-game; Crowd & Patent: Non-game > Game	None	Crowd & News: Game > Non-game; Student & News: Non-game > Game	Crowd & News: Game > Non-game

Table 27: Analysis of hypotheses for the acronym identification and resolution tasks

Metric	Main Effect	Interaction Effects
Acronym Identification		
Precision	Reject H_O	Reject H_O
Recall	Do not reject H_O	Reject H_O
F-score	Do not reject H_O	Reject H_O
Acronym Resolution		
Precision	Reject H_O	Do not reject H_O
Recall	Reject H_O	Reject H_O
F-score	Reject H_O	Reject H_O

5.10 Analysis

From the summary of findings (shown in Table 26) and the analysis of our hypotheses (shown in Table 27), we can make some observations. The game interface provides higher precision in the acronym identification task; but this advantage of the game interface does not carry over to the acronym resolution task. One possible reason may be due to the time limit in the game format this fast-past nature of a game interface lends itself to acronym identification, which depends more on recognition than on recall. Overall acronym identification is likely easier to accomplish than acronym resolution but identification tends to be tedious; therefore, gamifying this task may help to maintain a participant's concentration more than a non-game interface.

With recall, it is a mixed picture: the non-game interface provides higher recall in the acronym identification task for students evaluating News and the crowd evaluating Patents, but recall for the game interface was higher for the crowd evaluating the News collection. Again, this may point to the simpler acronyms being requested in the News collection, and the game helps keep the participant's attention.

For the acronym resolution task, the collection and participant are important factors: the News collection provides a higher recall than Patent collection and students provides a higher recall than the crowd. For recall, two of the three interaction effects observed with the acronym identification task also occur with the acronym resolution task: when students evaluate the News collection to resolve acronyms, they do better with a non-game interface; when the crowd evaluates the same News collection, they do better with a game interface. This may be a result of the instant feedback provided by the game interface, encouraging better performance from participants who enjoy challenges. Thus, the crowd's better recall performance with the game interface may have been a fortunate anomaly. Acronym resolution typically requires more careful consideration than acronym identification and the non-game interface may minimize many of the distractions that come with the game, such as performing the task quickly.

For acronym identification, the crowd obtains a higher F-score using the game interface in the News collection but using the non-game interface in the Patent collection. For acronym resolution, the News collection and the student participants obtain higher F-scores scores than their counterparts, but the same F-score advantage the game provides the crowd in the News collection for acronym identification also occurs with acronym resolution.

Table 28: Means and Standard Deviations comparing metrics for the human computation (HC) and algorithmic approaches.

	Precision		Recall		F-score	
Condition	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Acronym Identification						
News						
HC Approach (Mean)	0.975	0.039	0.606	0.129	0.739	0.064
Algorithmic Approach	0.855	0.178	0.892	0.206	0.873	0.194
Patent						
HC Approach (Mean)	0.985	0.037	0.640	0.149	0.764	0.142
Algorithmic Approach	0.813	0.193	0.891	0.199	0.850	0.196
Acronym Resolution						
News						
HC Approach (Mean)	0.968	0.027	0.498	0.069	0.655	0.064
Machine-based Approach	0.550	0.194	0.516	0.167	0.532	0.172
Patent						
HC Approach (Mean)	0.929	0.088	0.366	0.080	0.522	0.097
Algorithmic Approach	0.441	0.181	0.394	0.197	0.416	0.185

5.10.1 Algorithmic Approach vs. Human Computation Approach

One of our research questions is to examine the differences between the algorithmic approach and the human computation approaches we have empirically examined. We take the mean human computation score for each metric for each document in each collection

(n=1440) and compare it against the score obtained by the algorithm for each document (n=20). Table 28 illustrates the means and standard deviations for this comparison for each collection, and Table 29 provides the two-tailed t-tests for the group means for each metric for each collection.

Table 29: Two-tailed t-test results for the comparison of the mean human computation and algorithmic approaches for each metric for each collection

		Precision		Recall		F-score	
Condition	**df**	**T**	**p**	**t**	**p**	**t**	**p**
Acronym Identification							
News	1458	12.182	< 0.001	9.749	< 0.001	8.840	< 0.001
Patent	1458	17.826	< 0.001	7.444	< 0.001	2.674	0.032
Acronym Resolution							
News	1458	53.372	< 0.001	1.124	0.016	8.210	< 0.001
Patent	1458	24.127	< 0.001	1.506	0.132	4.772	< 0.001

From Tables 28 and 29, we can observe that for both collections for the acronym identification task, the mean human computation approach has higher precision but a lower recall and F-score than the Algorithmic approach. For the acronym resolution task, the mean human computation approach has a higher precision but lower recall for the News collection only. The F-score for acronym resolution is higher using the mean human computation approach in both collections.

We now assume that we have chosen the best overall strategy for each collection and use this knowledge as hindsight. Instead of using the mean human computation scores across all strategies, we consider the scores from the single best human computation approach for each collection and wish to evaluate this approach against an algorithm. For News, this is the non-game interface with student participants (n=360); for Patents, this is the non-game interface with crowd participants (n=360). Taking the averages again, we compare these two

human computation approaches with the Algorithmic approach (n=20) for each collection in Table 30. Table 31 provides the results from the two-tailed t-tests for the group means for each performance measure for each collection.

From Tables 30 and 31, the few advantages the Algorithmic approach had over the mean human computation approach are reduced. Compared with the best human computation approach in acronym identification, the Algorithmic approach no longer can claim a significantly better F-score for the Patent collection. For acronym resolution in both collections, the best human computation approach is now better across all three metrics, with the exception of recall for the Patent collection where there is now no significant difference between the human computation and algorithmic approaches.

Table 30: Means and standard deviations comparing metrics for the best human Computation (HC) approaches and the algorithmic approach.

Condition	Precision		Recall		F-score	
	Mean	SD	Mean	SD	Mean	SD
Acronym Identification						
News						
Best HC Approach for News	0.977	0.035	0.673	0.092	0.794	0.073
Algorithmic Approach	0.855	0.178	0.892	0.206	0.873	0.194
Patent						
Best HC Approach for Patents	0.988	0.015	0.709	0.129	0.818	0.109
Algorithmic Approach	0.813	0.193	0.891	0.199	0.850	0.196
Acronym Resolution						
News						
Best HC Approach for News	0.972	0.023	0.566	0.042	0.715	0.039
Algorithmic Approach	0.550	0.194	0.516	0.167	0.532	0.172
Patent						
Best HC Approach for Patents	0.929	0.090	0.398	0.092	0.554	0.111
Algorithmic Approach	0.441	0.181	0.394	0.197	0.416	0.185

Table 31: Two-tailed t-test t-values and p-values for the comparison of the best human computation (HC) approaches and algorithmic approach for each metric for each collection

Condition	df	Precision		Recall		F-score	
		t	P	t	p	t	p
Acronym Identification							
News	1458	10.116	< 0.001	3.924	0.001	4.124	< 0.001
Patent	1458	16.679	< 0.001	5.939	< 0.001	1.212	0.226
Acronym Resolution							
News	1458	37.541	< 0.001	1.124	0.016	14.712	< 0.001
Patent	1458	21.980	< 0.001	0.174	0.862	5.185	< 0.001

Table 32: Examination of the use of Common Knowledge (CK) in acronym resolution.

Factor	Precision		Recall		F-score	
	With CK	Without CK	With CK	Without CK	With CK	Without CK
Collection						
News	.9681	.5650	.4946	.3030	.6547	.3945
Patent	.9291	.4465	.3655	.2810	.5246	.3449
Participant						
Crowd	.9458	.4327	.4167	..2785	.5785	.5570
Student	.9513	.5788	.4644	.3155	.6241	.4084
Interface						
Game	.9460	.4810	.4302	.2890	.5914	.3611
Non-game	.9512	.4585	.4316	.2950	.5938	.3590

As discussed earlier, one of the strengths of using humans in information gathering tasks such as acronym resolution is the diversity of inputs. Indeed, the diversity that feeds consensus decisions has been shown to obtain more accurate results than could be obtained from a single participant, e.g., (Ashton, 1985; Michaelsen, Watson, & Black, 1989). This is particularly true for tasks requiring diverse knowledge like resolving acronyms. What is unclear is if humans are simply better at locating definitions or whether they are able to apply

external common knowledge (CK) to acronym resolution. Therefore, we examine how many of the definitions successfully resolved were determined by definitions in the text and how many were resolved using CK. To do so, we compare the mean human computation scores obtained with the scores that do not consider those acronyms where CK must be applied as determined by our human judges. This difference, illustrated in Table 32, allows us to observe the impact of using CK in acronym resolution.

From Table 32, we can observe the importance of human external knowledge in the acronym resolution process. We see that the use of external common knowledge is essential to obtaining a higher recall, precision and F-score and is the basis for the advantage of human-augmented computations in acronym resolution.

Finally, a key motivation behind this study is to examine the ability of humans to augment machine output in an effort to improve quality. Therefore, we are interested to examine how humans can assist with the acronym definition the algorithm incorrectly resolves or cannot find. We also examine the majority decision from the first three crowd participants to participate as well as the best three crowd participants for each collection based on the F-score.

5.10.2 Examining other methods to improve task quality

We have observed how human computation can improve acronym resolution over algorithmic approaches overall. Alonzo and Mizzaro (Omar Alonso & Mizzaro, 2012) observed an increase in relevance assessment quality using as few as three crowd participants and taking the majority (consensus) decision; however, Kamar *et. al.* (Kamar, Hacker, & Horvitz, 2012) found that accuracy observed their GalaxyZoo crowdsourcing task actually dips until at least 9 crowdworkers are used to make a majority decision. We look at first 3 individual crowd participants to participate from each collection and take the consensus decision of these participants; if at least 2 of the first 3 crowd participants could correctly resolve the acronym, we count it as success; if one or fewer resolve the acronym, we count it

as failure. This information, reported in Table 33, indicates a consensus decision with as few as 3 votes is as good as, or better than, the best individual scores in both collections.

Recent discussion in the crowdsourcing literature has evaluated the notion that a qualification or screening test would provide a more qualified pool of crowd participants. The qualification test requirement has been shown to be effective in previous studies involving the crowd (O. Alonso & R. Baeza-Yates, 2011; Carsten Eickhoff & de Vries, 2011; Heer & Bostock, 2010); however Ribeiro *et. al.* (Ribeiro, Florencio, & Nascimento, 2011) have found that qualification tests shrink the applicant pool without providing an appreciable improvement in quality. We wish to examine if providing the best qualified crowd participants in a task like acronym resolution can improve this quality. To accomplish this, we examine the performance of the top 3 participants and take the majority decision as we did with the first three crowd participants. These results appear in Table 34.

Table 33: Acronym recognition scores of the first 3 crowd participants to participate from each collection

Collection	Participant ID	Interface	Precision	Recall	F-score
News	10	Game	0.986	0.504	0.667
	1	Game	0.927	0.547	0.688
	31	Non-game	1.000	0.439	0.610
	Majority Decision		0.975	0.568	0.718
Patent	13	Game	1.000	0.397	0.569
	29	Non-game	0.957	0.468	0.629
	7	Game	0.984	0.433	0.601
	Majority Decision		0.985	0.482	0.647

Table 34: Acronym recognition scores of the 3 crowd participants from each collection with the best F-score, by collection.

Collection	Participant ID	Interface	Precision	Recall	F-score
News	12	Game	0.963	0.561	0.709
	15	Game	1.000	0.525	0.689
	16	Game	0.951	0.561	0.706
	Majority Decision		0.988	0.612	0.756
Patent	25	Non-game	0.972	0.489	0.651
	17	Game	1.000	0.496	0.664
	35	Non-game	1.000	0.539	0.700
	Majority Decision		0.988	0.568	0.721

The results observed in Table 34 indicate that using F-score and recall as performance measures, a consensus decision with as few as 3 votes is an improvement over the best individual scores in both collections, and indicates that restricting the crowd to the most qualified participants can provide better results such as acronym resolution, which require some external common knowledge.

5.10.3 Augmenting Algorithmic Approaches with Human Computation

Instead of comparing human computation methods with algorithmic approaches, we wish to observe how human computation methods can augment the algorithmic approach. We take the acronyms that the algorithm cannot resolve (i.e., provides an output of "undefined") and add the majority decisions from the two human computation groups (the first three participants or the best three participants) on these unresolved acronym definitions. We observe how this crowd augmented approach improves recall and precision for the algorithm alone. Table 35 indicates the improvement for each collection in precision, recall

and F-score using the first three participants (provided in Table 33) and the three best-performing participants (provided in Table 34) on acronym resolution.

From Table 35, we notice that a number of scores that include both the algorithm and the majority determination from either group of crowd participants increase the final precision and recall to a value considerably higher than the algorithmic approach alone. This indicates that we can potentially obtain quality improvements by implementing an approach using the majority decision from as few as three crowdworkers. This backs up our earlier hypothesis that augmenting algorithmic approaches can provide substantial value to tasks like acronym resolution. This also validates the value of the human-augmented approach from Criteria 6 of our framework and backs up other crowd assessment studies in the literature (O. Alonso & Mizzaro, 2009; Stefanie Nowak & Ruger, 2010).

Table 35: Improvement of precision and recall in acronym resolution using two human-augmented consensus approaches.

Collection	Performance Measure					
	Precision	**% Improv**	**Recall**	**% Improv**	**F-score**	**% Improv**
Algorithm Alone						
News	0.550	N/A	0.516	N/A	0.532	N/A
Patent	0.441	N/A	0.394	N/A	0.416	N/A
Algorithm + Consensus of First 3 Participants						
News	0.998	81.45	0.972	88.37	0.943	77.26
Patent	0.964	118.59	0.763	93.65	0.829	99.28
Algorithm + Consensus of 3 Best Performing Participants						
News	1.000	81.82	0.770	49.22	0.954	79.32
Patent	0.990	124.49	0.727	84.52	0.862	107.21

5.11 Conclusion

In this study, we have examined the performance of humans and algorithms to two tasks: the identification of acronyms and the resolution of acronyms. We use a rule-based algorithm that has successfully been applied to a collection of medical documents and apply it to two new collections, a collection of News articles and a collection of Patents. We examine factors that affect human computation: two modes of interaction (game and non-game interfaces) and two participant types (students and crowdworkers).

Through an examination of main and interaction effects on precision, recall, and F-score, we find that both human and algorithms identify and resolve acronyms in the News collection better than in the Patent collection. Students are better than the crowd at resolving the more difficult Patent collection acronyms, but there is no measurable difference between students and the crowd in the resolution of the easier News collection acronyms. Both participant types achieved better performance using the game interface to identify acronyms, but not in acronym resolution. Thus, game interfaces appear better suited to easier, yet mundane tasks, while non-game interfaces appear better suited for tasks that require more concentration, such as resolving more challenging acronyms in the Patent collection.

Compared with the algorithm we examined, humans have higher precision in identifying and resolving acronyms, while the algorithm is better at recall for these tasks. Humans are able to apply external common knowledge to the documents they examine, giving them a key advantage the algorithm does not have. The algorithm is better at applying rule sets to finding acronyms and their definitions. The best performance, however, is when we apply an augmented approach; that is, we use our algorithm to identify and resolve as many acronyms, then for those it cannot handle, we turn over to the humans to apply their external common knowledge. Performance increases substantially across all of our metrics when we use the augmented approach, in many cases doubling the performance.

CHAPTER 6

RELEVANCE ASSESSMENT STUDY

6.1 Background and Motivation

In the era of the Cranfield experiments of the 1960s (Cleverdon, 1967), often cited as the beginning of computer-based retrieval system evaluation, retrieval effectiveness of test databases was examined in controlled, laboratory-like settings. With smaller collections like Cranfield, exhaustive judgments of relevance for each query and document pair could be readily obtained. For larger modern collections, however, it is usual for relevance to be assessed only for a subset of all documents submitted for each query.

One standard approach to address this is *pooling*, a technique employed to evaluate relevance judgments to reduce human efforts. Pooling is used where relevance is assessed over a subset of the collection that is formed from the top "k" documents returned by a number of different IR systems (usually the systems being evaluated). For example, in TREC relevance assessments, each submitted run supplies the 1000 top-ranked documents for each topic. Of these, only the top 100 from each system are collected into a pool for human assessment. The evaluation is conducted with the assumption that all relevant documents are contained within the pool. This relevance assessment paradigm provides an abstraction of operational retrieval tasks in which a static set of relevance judgments are substituted with the complex interactions of a live (human) searcher.

The primary motivator for the use of pooling is to reduce the cost of relevance assessment. For example, if there were 100 runs, each which contribute 500 unique documents (after the removal of duplicates) to the number to be assessed, the cost would be overwhelming. If we were to assume a cost of $1.00 per document for manual assessment, the cost of the assessment would be $50,000, which is substantial. Therefore, a method that would obtain the same quality as pooling, yet reduce the number of documents needed for manual assessment, would provide a substantial benefit for relevance assessment. The recent

TREC Crowd campaigns have looked at this very issue, looking to see the best way in which the crowd can apply their collective relevance assessment abilities and maintain quality while simultaneously reducing assessment cost. The application of human computation methods to reduce assessment costs is the goal of this experiment as well.

One issue is to explore how the diverse backgrounds of the crowd assessors and the incentives of the crowdsourcing models might directly influence the trustworthiness and the quality of the resulting data. Several studies (e.g., (Berto, Mizzaro, & Robertson, 2013; Hosseini, Cox, Milić-Frayling, Kazai, & Vinay, 2012; R. McCreadie, et al., 2013) have independently examined crowdsourcing quality in relevance assessment and found that the quality can be obtained if the study is designed properly.

In our study, we compare the utility of two different algorithms, which are designed to reduce the number of non-relevant documents humans must assess. This extends some of the methods we applied in our submissions to TREC Crowd ’12 (C. Harris & P. Srinivasan, 2012), where we achieved the best overall scores across the performance measures examined. (Smucker et al, 2012). Unlike the methods used by many of the other participants in TREC Crowd’12, the methods we employ in this study are simpler, more scalable, easier to extend to other relevance assessment campaigns, and easier to implement. This study also examines the power of the crowd in order to assess relevance in two collections – one is a general News collection, another is a collection of Medical Documents. We examine how we can maintain quality while reducing cost in the evaluation of the query results against an information need, which is Step 10 of our IR model.

6.2 Contributions

In this study, we examine the use of machine and human computation methods to serve as an alternative to pooling. Unlike the methods conducted in our previous acronym study, in this study, we are only looking at a single mode of interaction (non-game interface). We use recall, precision, F-score, and LAM as our performance measures.

This experiment offers the following contributions. First, we examine a new approach to pooling that will reduce cost while still maintaining quality. Second, we examine if the use of human computation can be enhanced with an algorithmic approach for relevance judgments. Third, we apply our augmented human-machine algorithm methods to two datasets with very different characteristics.

6.3 Algorithms

In general, we select document and obtain judgments through crowdsourcing. We explore 2 document selection algorithms. Each operates on the merged set of documents retrieved by the different competing systems in TREC. We refer to this as the 'topic-document' set. For example, for topic *x*, system *y* submitted *z* unique documents. This *z* forms the 'topic-document' set for topic *x*.

6.3.1 Algorithm 1 – Merged ranking (non-clustered)

For each topic in our training set, we create a single ranked list. This list contained each document for that topic, ranked by $C(d)_W$ in decreasing order. The premise is that the relevant documents will have a high $C(d)_W$ and appear towards the top of our ranked list and the non-relevant documents will have a low $C(d)_W$ and appear towards the bottom.

Next, we determine an appropriate size for the batch of documents to send for crowd assessment. Through a training simulation used in our TREC 12 Crowd submission (C. G. Harris & P. Srinivasan, 2012), we empirically determine that the most appropriate batch size is 20. If a crowd participant judges all documents in a single batch as "not relevant", it marks the lower bound of the threshold between our potentially relevant and non-relevant document group. Therefore, we mark all the documents ranked below this relevance threshold as "not relevant". Using this approach on our TREC'12 Crowd simulation demonstrates the

potential merits with regard to time and cost: in that simulation, we are able to achieve a 90.1% recall by examining an average of only 21.5% of documents per topic.[33]

6.3.2 Algorithm 2-Merged ranking (clustered).

For Algorithm 3, we cluster the documents using k-means, which aims to partition documents for a given topic into k clusters in which each observation belongs to the cluster with the nearest mean. This results in a partitioning of the data space into Voronoi cells (Aurenhammer, 1991). We use k=11 for all clustering exercises.

We cluster only those documents that were part of the submitted runs for each topic. We cluster using the text in the title/headline and first 7 sentences of text (for documents belonging to the News collection), or title and the abstract (for documents belonging to the OHSUMED collection).

Once clusters are established, documents are ranked within each cluster, from highest to lowest $C(d)_W$. We begin with the cluster with the highest mean $C(d)_W$ score and ask crowd participants to evaluate documents in batches of 20 from that cluster in decreasing $C(d)_W$ rank order. As with the non-clustering algorithm, we use the crowd to mark the lower bound of the threshold between our potentially relevant and non-relevant document groups: if a batch of 20 comes back with all crowd-evaluated documents marked as “not relevant”, we mark all the remaining documents in that cluster that are ranked below this relevance threshold as “not relevant” and move to the cluster with the next-highest mean $C(d)_W$ score. Since we have a minimum of one batch from each cluster evaluated by the crowd, we have a minimum of 220 documents judged for relevance.

---

[33] In this simulation, we assume perfect crowd relevance assessment. In Run 2 of our TREC 12 Crowd TRAT submission, which we submitted to the crowd (and did not assume perfect relevance assessment), we used a batch size of 40, we obtained a 75.4% recall by crowd evaluation of only 17.2% of the overall document collection.

6.4 Hypotheses

We explore the following four hypotheses.

6.4.1 Algorithm Precision

$H_{0\text{-}1P}$: there are no main or interaction differences in the mean precision of relevance assessment due to the algorithm used.

$H_{A\text{-}1P}$: there are main or interaction differences in mean precision of relevance assessment due to the algorithm used.

6.4.2 Algorithm Recall

$H_{0\text{-}R}$: there are no main or interaction differences in recall of relevance assessment due to the algorithm used.

$H_{A\text{-}R}$: there are main or interaction differences in recall of relevance assessment due to the algorithm used.

6.4.3 Algorithm F-score

$H_{0\text{-}F}$: there are no main or interaction differences in the logistic average misclassification rate of relevance assessment due to the algorithm used.

$H_{A\text{-}F}$: there are main or interaction differences in the logistic average misclassification rate of relevance assessment due to the algorithm used.

6.4.4 Algorithm LAM

$H_{0\text{-}L}$: there are no main or interaction differences in the logistic average misclassification (LAM) rate of relevance assessment due to the algorithm used.

$H_{A\text{-}L}$: there are main or interaction differences in the logistic average misclassification (LAM) rate of relevance assessment due to the algorithm used.

6.5 Datasets and Topics

6.5.1 Selection of Topics

We select topics from two TREC test collections:

General Document Collection (News): These are taken from TREC-8 ad hoc task topics 401-443. This collection was discussed in Section 5.4.

Specialized Document Collection (OHSUMED): TREC-9 filtering track dataset – the OHSU-created test topics 1-43.

To obtain the topics for our evaluation, we obtain the list of documents submitted in TREC-9 OHSU filtering task for each of the 43 OHSUMED test topics (S. Robertson & Hull, 2000). Next, we select 3 topics from our News collection. We use topics 401-443, provided the TREC-8 ad hoc test collection (Voorhees & Harman, 1999).

The mean percentage of relevant documents for the entire collection was 1.29% for OHSUMED and 0.57% for News, a ratio of 2.26:1. Since we wish to compare our methods between collections, we identified the three comparable topics in each of the two collections in the following manner. First, we looked at the minimum and maximum percentage of relevant documents for each topic in the two collections. For the 43 News topics, the percentage of relevant documents per topic was between the range (0.03, 2.46). For the 43 OHSUMED topics, the percentages of relevant documents per topic fell between the range (0.04, 4.62). We ranked the OHSUMED topics by the percentage of relevant documents in the overlapping region (0.04, 2.46), and 38 of the 43 topics fall within this range. We then break this ranked list of 38 documents into three groups of roughly equal size (less than average % relevancy, average % relevancy, and greater than average % relevancy). We randomly select one topic from each of these three groups, which become our 3 OHSUMED topics. We then looked for 3 topics in the News collection that most closely matched the relevance percentages from the 3 randomly-selected OHSUMED topics. Selected were topics with characteristics mentioned in Tables 36 and 37.

Table 36: Characteristics of the selected topics from the OHSUMED collection

TopicID	Number of Submitted Docs	Number of Relevant Docs	Percentage Relevant
12	5291	7	0.132%
1	5784	44	0.761%
13	5841	77	1.318%

Table 37: Characteristics of the selected topics from the News collection

TopicID	Number of Submitted Docs	Number of Relevant Docs	Percentage Relevant
403	15636	21	0.134%
421	11090	83	0.748%
436	13940	180	1.291%

6.5.2 Selection of Documents

We use documents from two TREC test collections:

General Document Collection (News): Data collection used for the TREC-8 ad hoc task (TREC disks 4 and 5 less the Congressional Record) – which is a general News collection that has been used in many relevance assessment tasks. We use the test documents for the TREC-8 ad hoc task.

Specialized Document Collection (OHSUMED): TREC-9 filtering track dataset – a specific collection of OHSUMED (MESH) abstracts, titles, and MeSH (Medical Subject Heading) terms (Hersh, Buckley, Leone, & Hickam, 1994).

Both collections are created as follows. For each selected topic, all documents included in the submitted runs of past task participants are selected. This selected set of documents becomes our document collection for that topic. There were 129 and 53 submitted runs, respectively, for the News collection (TREC-8 ad hoc task) and OHSUMED (TREC-9 filtering task). Documents from the OHSUMED collection contain only the document abstracts whereas the documents from the News collection contain the entire document text. To compensate for this difference, we take the following approach. For

News collection documents, we use the headline and first seven sentences of text; for the OHSUMED documents, we use the headline and the full abstract. Statistics about the two collections are provided in Table 38.

Table 38: Text statistics from the documents in the News and OHSUMED collections

Text Statistic	OHSUMED		News	
	Mean	**SD**	**Mean**	**SD**
No. of sentences	6.97	1.04	6.99	0.04
No. of words	149.24	18.48	140.63	21.07
No. of complex words	32.04	3.96	24.90	10.63
Percent of complex words	22.19%	0.02	17.71%	0.03
Average words per sentence	21.13	1.39	20.12	0.48
Average syllables per word	1.76	0.09	1.68	0.11

6.6 Gold standard

For the experiment using the News collection, we obtain our gold standard from the binary relevance assessments provided from the TREC-8 ad hoc tasks. We randomly select 3 test topics from topics 401-443 provided the TREC-8 ad hoc test collection (Voorhees & Harman, 1999).

For the experiment against the OHSUMED collection, we obtain our gold standard from the OHSUMED relevance assessments provided in the TREC-9 filtering task (Robertson & Hull, 2000). Since the documents in this collection are given one of three relevance states by the OHSUMED assessors (0 = “not relevant”, 1 = “partially relevant”, 2 = “definitely relevant”), we follow the liberal approach used by the OHSUMED assessors (Hersh, et al., 1994) as well as suggested in (Omar Alonso & Mizzaro, 2012) and group the “partially relevant” and “definitely relevant” documents as “relevant”.

6.7 Experimental Setup

6.7.1 Determining an appropriate value of k for k-means clustering

We rank the topics from lowest percentage of relevant documents (as determined by the gold standard) to the highest percentage. We then break this into three groups of roughly equal size. We perform k-means clustering on the title/heading and first seven sentences of document text (for News) and title/heading and full abstract (for OHSUMED). As indicated in several studies, e.g., (Bradley & Fayyad, 1998; Fraley & Raftery, 1998; Pham, Dimov, & Nguyen, 2005), the expectation–maximization (EM) approach (Dempster, Laird, & Rubin, 1977) is recognized as the most common method for selecting k for k-means clustering; we take this approach to set the value for k in our study as well.

To exploit the use of clustering to help reduce the number of documents to assess, we desire a value of k that provides the greatest variance. A greater intra-cluster variance (i.e., less uniformity) indicates some clusters that have a larger percentage of relevant documents while other clusters will have a smaller percentage. If we can select those clusters that contain a high percentage of relevant documents, we can reduce the number of documents necessary for assessment.

Consistent with other methods to establish an initial value of k for k-means (e.g., (Duda & Hart, 1973; Meilă & Heckerman, 1998)e take the following approach. For each of these randomly-selected topics, we use k=5 and calculate the variance in the number of relevant documents that appear in each cluster. We repeat this exercise increasing k in Steps of 3 until we obtain a k=20 (e.g., k values of 8, 11, 14, 17, and 20) and examine the variance in the relevant document percentages across each cluster. We then evaluate the values of k on each side of the high value, and repeat this once more until we obtain a maximum variance. For the low, medium, and high topics in OHSUMED, we obtain values for k of 11, 12, and 12 respectively. For News, the values for k were 10, 9, and 13 respectively.

Tables 39 and 40 contain the training run variances for values of k for these training runs for three representative OHSUMED and News topics, respectively. As a result, we set k=11 for both collections in our experiment.

Table 39: Variance for different values of k for three OHSUMED topics.

Topic ID	Rel %	5	8	9	10	11	12	13	14	17	20
10	Low	0.0105	0.0103	0.0106	0.0128	**0.0136**	0.0132	0.0125	0.0118	0.0112	0.0116
2	Mid	0.0024	0.0018	0.0034	0.0049	0.0084	**0.0088**	0.0070	0.0069	0.0054	0.0058
33	High	0.0165	0.0174	0.0196	0.0218	0.0247	**0.0263**	0.0246	0.0219	0.0196	0.0208

Table 40: Variance for different values of k for three News topics.

TopicID	Rel %	5	8	9	10	11	12	13	14	17	20
406	Low	0.0035	0.0062	0.0073	**0.0078**	0.0064	0.0059	0.0051	0.0042	0.0039	0.0029
413	Mid	0.0056	0.0066	**0.0085**	0.0071	0.0063	0.0054	0.0051	0.0037	0.0032	0.0027
404	High	0.0031	0.0053	0.0060	0.0067	0.0076	0.0082	**0.0089**	0.0084	0.0077	0.0063

6.7.2 Creating our Document Ranking

The use of previous run information demonstrated its effectiveness in our TREC Crowd'12 TRAT run submissions (C. Harris & P. Srinivasan, 2012). This approach can be explained as follows. For each selected topic, we gathered the submitted runs from the appropriate TREC tasks. These submission files contain submitted runs using a variety of methods and sources; however each contains the topic ID, retrieved document name and a binary relevance score. Table 41 contains the number of documents and submitted runs from each collection across all topics for each collection.

Table 41: Characteristics of the submitted runs for each collection.

Document Collection	TREC Task	Number of Topics	Number of Submitted Runs	Number of Documents
News	TREC-8 ad hoc	3	129	40666
OHSUMED	TREC-9 filtering	3	75	16916

6.7.3 Determining weighted counts for each document

Using the ranked lists of documents from TREC participants, we compute two scores for each topic: (1) A *simple count*, C_S, indicates the count of submitted runs that included a given document, (2) A *Borda count*, C_B, takes into account the rank in each submitted run for a given document. This represents an approach similar to the one used in (Almquist, Mejova, Ha-Thuc, & Srinivasan, 2008). This Borda count is calculated as *(n-r)*, where *n* is the number of documents retrieved for a topic in a single submission file, and *r* represents the document's rank within the list (i.e., the top-ranking document in a list of 1000 documents will receive a score of (1000-1) = 999). We then sum the Borda count for all available submissions for that topic. The TREC campaigns allowed a maximum of 1000 submitted documents per run per topic, so for each of our submitted runs for that collection, the Borda count is in the range (0, 999).

We use both counts since they represent different properties of each training document. C_S measures the number of submissions that include that document for a topic, but does not consider its rank; C_B examines the documents rank but does not consider how many of the submitted runs the document appears. For example, for a given topic in the News collection, if a document exists in all 129 submitted runs, it would receive a C_S of 129. However, if that document was ranked at the bottom of each list, the document is not likely to be relevant. Conversely, if a document was listed in only 10 of the 129 submitted runs, but ranked at or near the top of each, C_B would be relatively high. The *count ratio coefficient*,α, represented by a value in the range (0,1), is the relative balance between these two counts for a data collection. Using these two counts (C_S and C_B) and applying the count

ratio coefficient, α, we calculate a *weighted rank coefficient*, $C(d)_W$, for each document using these two separate counts for each individual document, *d*. A document will have a different weighted rank coefficient for each topic examined.

$$C(d)_W = \alpha\, C(d)_S + (1-\alpha)\, C(d)_B$$

A merged listing of documents was created ranked by $C(d)_W$, from highest to lowest for each topic.

We then experimented with various values of α, from 0 to 1, in increments of 0.05. A *relevant document score at* ~, S_α, was determined for each topic:

$$S_\alpha = \frac{\sum_{n=1}^{d} \mathrm{rel}(n) * C(n)_{W_\alpha}}{\sum_{n=1}^{d} \mathrm{rel}(n)}$$

Where *rel(n)* is the binary relevance for document *n* and $C(n)_{W_\alpha}$ is the weighted sum for document *n* for a given α for one topic. S_α indicates the weighted rank of all relevant documents for a single topic for a given α; we obtain the average S_α across all ten of our training topics. If we rank our list by $C(d)_W$ in decreasing order and the resulting S_α is large (i.e., documents appearing at the top are relevant), it indicates the selected α bunches the relevant documents closer to the top of our list. Empirically, we determined that α = 0.8 provided the highest S_α across all training set topics. We therefore use this value for calculating our *document score*.

6.8 Interfaces

The interfaces used are similar to those used in TREC Crowd '12 TRAT task (C. Harris & P. Srinivasan, 2012). Using a non-game web interface, the user is provided with the title and first 7 sentences of text for documents belonging (for the News collection), or title and the abstract for documents (for the OHSUMED collection) and asked to assess binary relevance (relevant/not relevant) on that document to a single provided topic.

6.9 Participants

Only crowd participants recruited from Amazon Mechanical Turk are used in this task. These crowd participants are assigned randomly to either the non-clustering or clustering algorithm and are compensated $0.60 per batch of 20 documents assessed. For each algorithm, we use 24 participants; each participant evaluates 3 topics from a single collection. Table 42 indicates the assignment of topics and participants by algorithm and collection.

Participants evaluate batches of documents for a single topic until they determine the relevance threshold for that topic. Participants who do not complete the full assessment of 3 topics have their assessments removed and the task is made available for other crowd participants.

Table 42: Assignment of topics by algorithm and collection.

Collection	Algorithm	Number of Topics	Number of Participants
News	Non-clustering	3	24
	Clustering	3	24
OHSUMED	Non-clustering	3	24
	Clustering	3	24

6.10 Results

Table 43 provides the means and standard deviations across our four dependent variables.

Table 43: Means and standard deviations for the dependent variables for comparison of algorithm and collection in relevance assessment (n = 96)

	Relevance Assessment								
	Precision		Recall		F-score		LAM		
Condition	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**N**
Algorithm type									
Non-clustering	0.610	0.089	0.543	0.102	0.551	0.101	0.043	0.007	48
Clustering	0.323	0.040	0.740	0.055	0.430	0.044	0.062	0.004	48
Collection type									
News	0.451	0.137	0.584	0.137	0.449	0050	0.056	0.007	48
OHSUMED	0.482	0.180	0.699	0.087	0.532	0.117	0.050	0.013	48
Algorithm × Collection									
Non-clustering, News	0.572	0.078	0.452	0.029	0.467	0.049	0.049	0.005	24
Non-clustering, OHSUMED	0.648	0.084	0.635	0.054	0.635	0.062	0.062	0.004	24
Clustering, News	0.331	0.043	0.717	0.035	0.431	0.045	0.037	0.003	24
Clustering , OHSUMED	0.315	0.036	0.764	0.062	0.430	0.044	0.062	0.005	24

6.10.1 Relevance Assessment Precision

Table 44: ANOVA results for precision of relevance assessment

	Precision				
Effect	**SS**	**df**	**Mean Square**	**F**	**p-value**
Algorithm	1.973	1	1.973	487.285	0.000
Collection	0.022	1	0.022	5.513	0.021
Algorithm x Collection	0.051	1	0.051	12.472	0.001
Error	0.373	92	0.004		

Table 44 provides the ANOVA results for precision for the relevance assessment task. ANOVA results for precision of relevance assessment were statistically significant ($p < 0.05$) for algorithm, with non-clustering algorithms scoring better than clustering algorithms, $F(1,92) = 487.285$, $p<0.001$, for collection, with OHSUMED collection receiving higher

precision than the News collection, F(1,92)=5.1513, p=0.021 and for the algorithm x collection interaction, F(1,92)=12.472, p=0.001. This two-way interaction is graphically depicted in Figure 11.

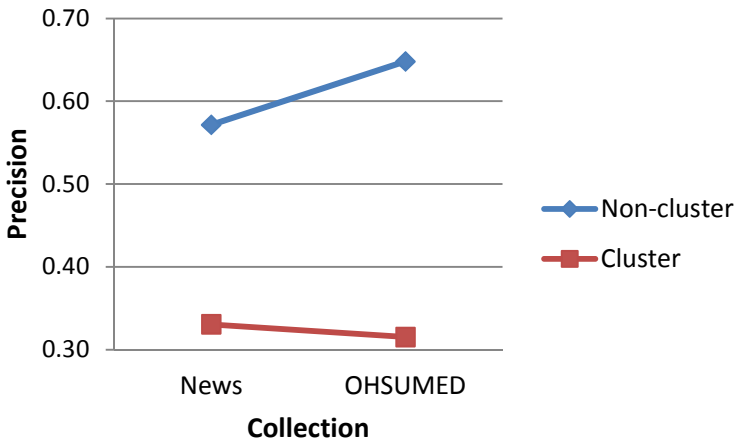

Figure 11: Interaction between algorithm (lines) and collection (x-axis) for precision in relevance assessment.

Simple effect follow-up tests for the two-way interaction revealed that precision was higher for the non-clustering algorithm than for the clustering algorithm in both the News collection (M_{NC}=0.572 M_C= 0.331; M_{Diff} = 0.241; t(92)=4.172; p<0.001; d=1.20) and the OHSUMED collection (M_{NC}=0.648 M_C= 0.315; M_{Diff} = 0.333; t(92)=5.761; p<0.001; d=1.66)

6.10.2 Relevance Assessment Recall

Table 45 provides the ANOVA results for recall for the relevance assessment task. ANOVA results for recall of relevance assessment were statistically significant ($p < 0.05$) for algorithm, with the clustering algorithm outperforming the non-clustering algorithm, F(1,92)= 416.222, p<0.001, for collection, with the OHSUMED collection providing better recall than the News collection, F(1,92)=141.348, p<0.001and for the algorithm x collection interaction, F(1,92)=49.112, p<0.001. This two-way interaction is graphically depicted in Figure 12.

Table 45: ANOVA results for recall of relevance assessment

Effect	Recall				
	SS	df	Mean Square	F	p-value
Algorithm	0.928	1	0.928	416.222	0.000
Collection	0.315	1	0.315	141.348	0.000
Algorithm x Collection	0.109	1	0.109	49.112	0.000
Error	0.205	92	0.002		

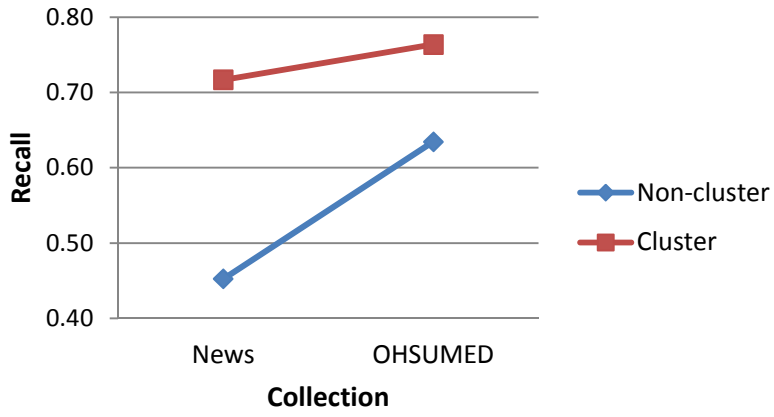

Figure 12: Interaction between algorithm (lines) and collection (x-axis) for recall in relevance assessment.

Simple effect follow-up tests for the two-way interaction revealed that recall was higher for the clustering algorithm than for the non-clustering algorithm in both the News collection (M_C=0.717 M_{NC}= 0.452; M_{Diff} = 0.241; t(92)=6.471; p<0.001; d=1.87) and the OHSUMED collection (M_C=0.648 M_{NC}= 0.315; M_{Diff} = 0.333; t(92)=3.162; p=0.002; d=0.91) .

6.10.3 Relevance Assessment F-score

Table 46: ANOVA results for F-score of relevance assessment

Effect	F-score				
	SS	df	Mean Square	F	p-value
Algorithm	0.349	1	0.349	92.300	0.000
Collection	0.168	1	0.168	65.730	0.000
Algorithm x Collection	0.171	1	0.171	66.781	0.000
Error	0.235	92	0.003		

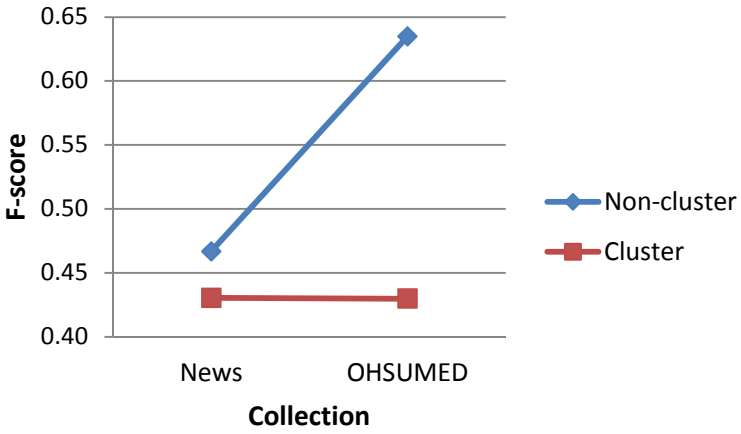

Figure 13: Interaction between algorithm (lines) and collection (x-axis) for F-score in relevance assessment.

Table 46 provides the ANOVA results for F-score for the relevance assessment task. ANOVA results for F-score of relevance assessment were statistically significant ($p < 0.05$) for algorithm, with the non-clustering algorithm outperforming the clustering algorithm, $F(1,92)=92.3$, $p<0.001$, for collection, with the OHSUMED collection obtaining a higher F-score than the News collection, $F(1,92)=65.73$, $p<0.001$ and for the algorithm x collection

interaction, F(1,92)=66.78, p<0.001. This two-way interaction is graphically depicted in Figure 13.

Simple effect follow-up tests for the two-way interaction revealed that F-score was higher for the non-clustering algorithm than for the clustering algorithm in both the News collection (M_{NC}=0.467 M_C= 0.431; M_{Diff} = 0.036; t(92)=2.287; p=0.024; d=0.66) and the OHSUMED collection (M_{NC}=0.635 M_C= 0.430; M_{Diff} = 0.205; t(92)=12.96; p<0.001; d=3.47)

6.10.4 Relevance Assessment LAM

Table 47: ANOVA results for LAM of relevance assessment

	LAM				
Effect	**SS**	**df**	**Mean Square**	**F**	**p-value**
Algorithm	0.009	1	0.009	428.432	0.000
Collection	0.001	1	0.001	41.175	0.000
Algorithm x Collection	0.001	1	0.001	44.851	0.000
Error	0.279	92	0.001		

Table 47 provides the ANOVA results for LAM for the relevance assessment task. ANOVA results for LAM of relevance assessment were statistically significant ($p < 0.05$) for algorithm, with non-clustering outperforming clustering, F(1,92)=428.432, p<0.001, for collection, with OHSUMED outperforming News, F(1,92)=41.175, p<0.001 and for the algorithm x collection interaction, F(1,92)=44.851, p<0.001. This two-way interaction is graphically depicted in Figure 14.

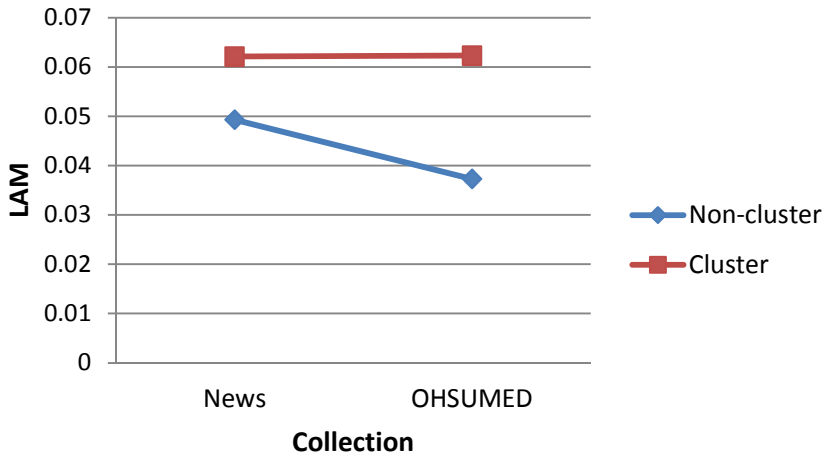

Figure 14: Interaction between algorithm (lines) and collection (x-axis) for LAM in relevance assessment.

For LAM, which is a misclassification rate, a lower number is considered a better result. Simple effect follow-up tests for the two-way interaction revealed that LAM was lower (better) for the non-clustering algorithm than for the clustering algorithm in the OHSUMED collection (M_{NC}=0.037 M_C= 0.062; M_{Diff} = 0.025; t(92)=2.739; p=0.007; d=0.791).

6.10.5 Summary of Findings and Evaluation of Hypotheses

Table 48 provides a summary of our findings for the Acronym Identification and Resolution tasks. From Table 48, we can assess our hypotheses. This assessment is provided in Table 49.

Table 48: Summary of findings for relevance assessment

Effect Type	Relevance Assessment			
	Precision	**Recall**	**F-score**	**LAM**
Main	Algorithm: Non-cluster > Cluster; Collection: OHSUMED > News	Algorithm: Cluster > Non-cluster; Collection: OHSUMED > News	Algorithm: Non-cluster > Cluster; Collection: OHSUMED > News	Algorithm: Non-cluster > Cluster; Collection: OHSUMED > News
Interaction	News: Non-cluster > Cluster; OHSUMED: Non-cluster > Cluster	News: Cluster > Non-cluster; OHSUMED: Cluster > Non-cluster	News: Non-cluster > Cluster; OHSUMED: Non-cluster > Cluster	OHSUMED: Non-cluster > Cluster

Table 49: Analysis of hypotheses for relevance assessment

Metric	Main Effect	Interaction Effects
Relevance Assessment		
Precision	Reject H_O	Reject H_O
Recall	Reject H_O	Reject H_O
F-Measure	Reject H_O	Reject H_O
LAM	Reject H_O	Reject H_O

6.11 Analysis

From the summary of findings (shown in Table 48) and the analysis of our hypotheses (shown in Table 49), we make some observations. We find that for both collections, the algorithm used affects all four performance measures, with the non-clustering algorithm providing better precision, F-score, and LAM, while the clustering algorithm

provides better recall. Likewise, we find that for relevance assessment, the OHSUMED collection outperforms the News collection for each of our four performance measures.

The difference between the non-clustering and clustering algorithm is that the former considers relies on the weighted count ranking, while the latter considers both the weighted count ranking and the k-means clustering method to determine the order in which documents are presented to participants. Since similar documents are clustered together, it is not surprising that recall is improved with the clustering algorithm; however, this comes at the expense of precision. Non-relevant yet similar documents also seem to be clustered together. When the two measures are contrasted in the F-score, the simpler non-clustering algorithm scores beats out the more complex clustering algorithm.

LAM, also known as LAM%, is a widely-applied metric used in evaluating spam filtering performance (Qi, Yang, He, & Li, 2010) and is also used in the TREC Crowdsourcing track It is designed to impose no *a priori* relative importance on those documents determined to be false positives or false negatives, yet equally rewards an improvement in the odds of either. A lower misclassification rate is a therefore a better result. In TREC Crowd'12 (C. Harris & P. Srinivasan, 2012), the clustering method we used, similar to the one used in this study, obtained the lowest LAM of all 33 submitted runs (Smucker, et al.). Our methods here obtained a LAM rate that is comparable to our best run in TREC Crowd'12.

The non-clustering algorithm's better precision, F-measure and LAM results have another possible explanation. The algorithm ranks documents based only on the weighted count ranking, which is derived from multiple run submissions. Therefore, we are able to observe the merits of using the diversity of run submissions to rank and documents, much as we use the diversity of the crowd to provide better relevance assessment results for each information need.

Clustering provided better recall in both collections. This higher recall may be due to the increased diversity that the clustering method provides. Since we are clustering based on

the text in each document, then ranking documents within each cluster, this allows the documents which are ranked in the highest 20 documents within each cluster to be guaranteed to be reviewed by the crowd. This allows a greater number of the documents deemed relevant by the TREC assessors to be evaluated, including many documents that would appear too far down the ranked list provided by the non-clustering algorithm.

One limiting factor of our methods is that we did not take advantage of all of the features provided for each document. With the News collection, for example, we limited our document to the first seven sentences of text in order to make it similar in size to the OHSUMED collection and avoid bias between our collections; however, the text containing data relevant to the information need may appear outside of these initial seven sentences in the News collection text. Likewise, to maintain collection compatibility, the OHSUMED collection did not include the Medical Subject Heading (MeSH) terms. We suspect would further improve our relevance assessment results with the OHSUMED collection.

6.11.1 Examination of Retrieval Efficiency

We wanted to compare our method with the pooling method; that is, how effective was our method at identifying the relevant documents (as determined by the TREC gold standard). We examine this by looking at the number of batches of 20 documents examined by the crowd for each combination of algorithm and collection. This information is presented below in Table 50.

Table 50: Comparison of results between TREC assessors and our methods, by topic and collection.

Collection	Topic ID	# Unique Docs	TREC Assessors			Non-clustering			Clustering		
			# Eval	# Rel	# NR	# Eval	# Rel	# NR	# Eval	# Rel	# NR
OHSUMED	1	5784	5784 (100%)	44	5740	180 (3.1%)	36 (81.8%)	144	640 (11.1%)	36 (81.8%)	604
	12	5291	5291 (100%)	7	5284	60 (1.1%)	4 (57.1%)	56	440 (8.3%)	5 (71.4%)	435
	13	5841	5841 (100%)	77	5764	300 (5.1%)	50 (64.9%)	250	800 (13.7%)	61 (79.2%)	739
News	403	15636	1046 (6.7%)	21	1025	100 (0.6%)	20 (95.2%)	80	720 (4.6%)	20 (95.2%)	700
	421	11090	1763 (15.9%)	83	1680	300 (2.7%)	20 (24.1%)	280	1040 (9.4%)	53 (63.9%)	987
	436	13940	1949 (13.4%)	180	1769	280 (2.0%)	38 (21.1%)	242	1240 (8.9%)	109 (60.6%)	1131

From Table 50, we can observe that the number of batches evaluated, relative to the number of possible batches, is small. The number of batches evaluated by participants using the clustering algorithm is greater than the number of batches evaluated by participants using the non-clustering algorithm. This is because our algorithm presents participants with at least one batch from each cluster.

From Table 50 we can also see that the number of documents in the News collections is much larger than that in the OHSUMED collection. The number of documents marked relevant by the TREC assessors in each collection varies, but is considerably greater than the number evaluated by either of our algorithms. The number of relevant documents found by each algorithm does not differ for topics with only a few relevant documents (OHSUMED Topic IDs 1 and 12; News Topic ID 403), but it does differ for topics with a large number of relevant documents. A second point that may have worked against the clustering method is that each cluster had, at a minimum, a batch of 20 documents judged. Retrospectively

speaking, we should have gauged the relevance of each cluster and eliminated those that were unlikely to contain any relevant documents.

It is somewhat surprising that the OHSUMED collections outperformed the News collections in each of our performance measures. This may be due to the use of the concise focus of using the abstract text in OHSUMED instead of the first seven sentences the News document. It could also be due to the broad nature of the topics to be evaluated for the News collection, coupled by more rules imposed on the participant to determine a document's relevancy. The rules to ascertain relevance used by the TREC assessors in the OHSUMED collection had far fewer limitations or restrictions on relevancy as compared with those used by the TREC assessors for the News collection.

As Table 50 illustrates, the non-clustering algorithm's efficiency drops more rapidly than the clustering algorithm in collections that contain many relevant documents. The efficiency of our algorithms could be modified to identify a larger set of relevant documents by employing methods such as using a larger batch size, using a majority (consensus) determination method of voting on document relevance, using all the text within each document, or making use of other available document features.

From Table 50, we can also observe the efficiency of each algorithm for each topic. On average, participants using the non-clustering algorithm only evaluated 3.1% of the OHSUMED documents, but were able to find (and properly assess) 68.0% of the relevant documents; participants using the clustering algorithm evaluated 11.0% of the OHSUMED documents, but were able to find (and properly assess) 77.5% of the relevant documents. For the News collection, the difference is greater: participants using the non-clustering algorithm only evaluated 1.8% of the entire collection but were able to find (and properly assess) 46.8%; for participants using the clustering algorithm, they evaluated 7.6% of the collection, but were able to find (and properly assess) 73.2% of the relevant documents. For the three topics in the News collection (which used pooling) 12.0% of the documents were judged by TREC assessors. Thus, as a proxy for pooling, the clustering method provides an efficient

coverage of the relevant documents at a fraction of the number of documents used in the pooling process.

These two algorithms' ability to work in collections with only a few relevant documents is particularly noteworthy – for example, the non-clustering algorithm only evaluated a mere 0.6% of the News collection for topic 403, but was able to find (and properly assess) 95.2% of the relevant documents.

6.11.2 Evaluating the Efficiency of our Methods

Table 51 presents each method by collection and topic ID, as evaluated by precision, recall, and F-score. This illustrates how each of these three approaches compares in terms of overall efficiency.

Table 51: The overall effectiveness of each method by precision, recall, and F-score performance measures, by topic and collection

Collection	Topic ID	TREC Assessors			Non-clustering			Clustering		
		Prec	Rec	F	Prec	Rec	F	Prec	Rec	F
OHSUMED	1	0.008	1.000	0.016	0.200	0.818	0.321	0.060	0.818	0.112
	12	0.001	1.000	0.003	0.070	0.571	0.125	0.010	0.714	0.020
	13	0.013	1.000	0.026	0.170	0.649	0.269	0.080	0.792	0.145
News	403	0.020	1.000	0.039	0.200	0.952	0.331	0.030	0.952	0.058
	421	0.047	1.000	0.090	0.070	0.241	0.108	0.050	0.639	0.093
	436	0.092	1.000	0.168	0.140	0.211	0.168	0.090	0.606	0.157

When evaluating the necessary costs for relevance assessment, the benefits of our method become more evident. Evaluation costs for relevance assessment are one of biggest constraints. Examining Table 51, we can examine the effectiveness of each method. For the OHSUMED collection, the average F-score for the non-clustering and clustering methods is 16.44 and 6.36 times that of the TREC pooling method; for the News collection, we achieve

average F-scores that are 2.04 and 1.03 times that of the TREC pooling method. Assuming the costs for TREC assessors and our methods are equal, the effectiveness of our methods is equal to or greater than the pooling method used by TREC. In reality, however, the costs per assessment for our crowd-based methods are substantially less. In the overview document for the 2007 TREC Legal track, one of the few published articles in which TREC relevance assessment costs are indicated, human assessors evaluated an average 20 documents per hour. The relevance assessment cost for that task was estimated by the authors at $150 an hour, or $7.50 per document (Tomlinson, Oard, Baron, & Thompson, 2007). Even if assessment costs for our task were only $2 per document, the cost of the TREC pooling process for the 3 topics in the News collection would be $9,516. In comparison, the cost for our study was $0.22 per batch of 20 documents, including Amazon Mechanical Turk overhead fees, which comes to slightly more than $0.01 per document. The cost for all 24 participants to evaluate the same 3 News topics was $792.00 for the clustering algorithm and $179.52 for the non-clustering algorithm. Using only 3 crowdworkers and taking the majority decision, as discussed in Alonso and Mizzaro (O. Alonso & Mizzaro, 2009; Omar Alonso & Mizzaro, 2012), we can reduce these costs by a further 87.5%.

6.11.3 The Effects of Reducing the Number of Clusters Evaluated

In our clustering algorithm, we had participants evaluate at least one batch from each of the 11 clusters. This may have limited the clustering method's performance when compared with the non-clustering algorithm. If we can establish a threshold on each cluster, and only allow the evaluation of clusters that meet or exceed that threshold, we can improve the performance overall.

For each topic in each collection, we rank the 11 clusters by their mean weighted score. We remove clusters from lowest score to highest score and evaluate precision, recall, and F-score for each collection. This is shown in Figure 15.

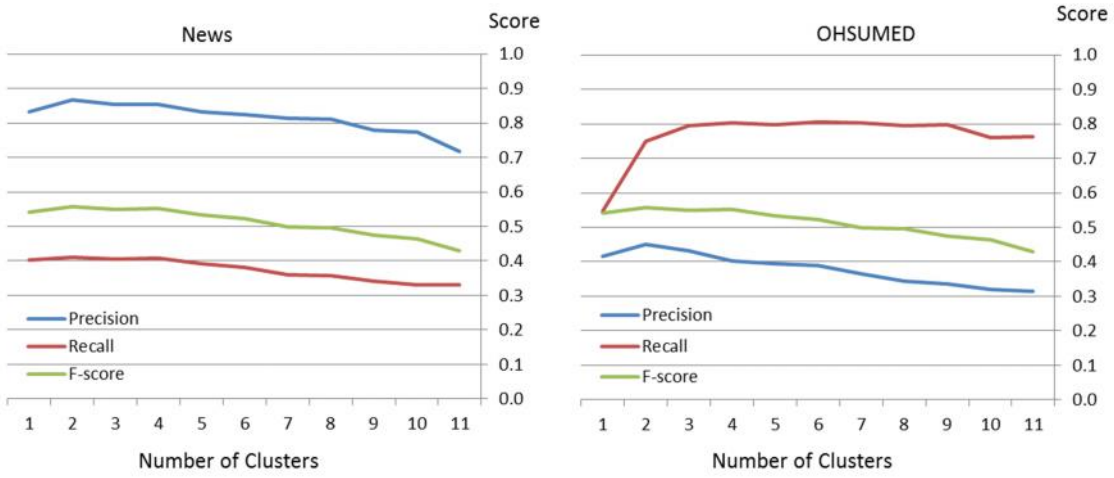

Figure 15: Measures of performance as the number of clusters are increased for News (left) and OHSUMED (right)

From Figure 15, in each graph, we observe that when all 11 clusters are evaluated, our F-score (middle line, in green) is at its lowest, indicating that the evaluation of all clusters is sub-optimal. In both News and OHSUMED collections, the best F-score is obtained when we evaluate only the first two clusters and ignore the rest. In Table 52, we examine the mean weighted score and the overall mean for each collection. Conservatively, if we eliminate all those clusters with a mean weighted score (wtscore) that is below the overall mean for the collection, we can reduce the number of clusters we evaluate and increase our overall performance. This threshold would have the crowd evaluate only the 3 highest-ranking clusters in the News collection and the 5 highest-ranking clusters in the OHSUMED collection, which are italicized in Table 52.

Table 52: Ranking of clusters by mean weighted score, by collection.

Cluster	News		OHSUMED	
	Mean wtscore	**% of Mean wtscore**	**Mean wtscore**	**% of Mean wtscore**
1	2207.44	*1.87*	903.40	*1.21*
2	1518.54	*1.29*	811.07	*1.09*
3	1359.21	*1.15*	786.29	*1.05*
4	1124.90	0.95	779.49	*1.04*
5	1049.56	0.89	752.92	*1.01*
6	1054.39	0.89	733.93	0.98
7	1052.30	0.89	717.68	0.96
8	989.07	0.84	703.83	0.94
9	1055.80	0.89	697.10	0.93
10	954.69	0.81	686.78	0.92
11	869.60	0.74	632.48	0.85
Collection Mean	1181.27	1.00	746.37	1.00

6.11.4 Detecting potentially relevant documents

A key strength of our approach is that it permits discovery of relevant documents that might have been overlooked by the pooling process. In the News collection, our participants were able to assess documents that were not a part of the TREC assessor pool. Table 53 provides a matrix that shows how our approaches' assessments compare with those made by the TREC Assessors. To avoid a small number of crowd participants skewing the results using our methods, a minimum threshold agreement of 25% (e.g., at least 12 of the 48 assessors) had to have evaluated that document to mark it as relevant or non-relevant.

From Table 53, the documents that are most interesting to us are those where at least 25% of the crowd participants have evaluated a document, but a majority of them disagree with the relevance decision made by the TREC assessors. The document IDs for the News and OHSUMED collections appear in Tables F-1 through F-5 in Appendix F.

Table 53: Document evaluation comparison matrix of our methods with the determination of the TREC assessors for the News collection (top) and OHSUMED collection (bottom)

News Collection

		Determination by Our Methods		
		Not Evaluated	Non-Relevant	Relevant
TREC Assessor Determination	Not Evaluated	34725	1174	9
	Non-Relevant	2789	1681	4
	Relevant	72	10	202

OHSUMED Collection

		Determination by Our Methods		
		Not Evaluated	Non-Relevant	Relevant
TREC Assessor Determination	Non-Relevant	14599	2188	1
	Relevant	17	3	108

6.11.5 Comparing Algorithms on a Larger Set of Topics

Next, we wish to examine the performance of each of our two strategies against a larger set of topics in a single collection. This deeper examination permits us to validate our earlier findings across a wider range of topics. In a second phase, we ask participants to evaluate 20 topics from the News collection, three topics being the same as we examined earlier in this study (topics 403, 421, and 436), along with 17 additional topics that were randomly selected from the same TREC campaign. In contrast with Phase 1, where only the first 7 sentences of text are provided (in order to keep the two collections similar), each participant in Phase 2 is given with the entire text of the news article for their assessment. In TREC pooling approach, the entire document is considered, so the Phase 2 comparison is a better indicator of performance between our algorithmic approaches and the pooling

approach. Also, in this phase we use a more realistic approach of having 3 judgments for each topic.

Similar to the earlier phase (Phase 1), all participants in this new phase (Phase 2) were required to evaluate 3 separate topics for their assessments to be considered in this study. Each document was evaluated for relevance by 3 different participants. The mean values of each algorithm for each performance measure are provided in Table 54 (bolded numbers imply the best performance for each measure).

Table 54: Mean values for each performance measure

Algorithm	**Precision**	**Recall**	**F-score**	**LAM**
Non-cluster	**0.8826**	0.6282	0.7270	**0.0499**
Cluster	0.8174	**0.7645**	**0.7861**	0.0524

Wilcoxon Signed-Rank tests were run to determine if there were differences between our algorithms for each of our four performance measures. For precision, the non-clustering algorithm (Median = 0.8956) significantly outperformed the clustering algorithm (Median = 0.8215), $z = -2.837$, $p < .0005$. For recall, the clustering algorithm (Median = 0.6517) significantly outperformed the non-clustering algorithm (Median = 0.8013), $z = 3.724$, $p < .0005$. For F-score, the clustering algorithm (Median = 0.8112) significantly outperformed the non-clustering algorithm (Median = 0.7651), $z = 3.061$, $p < .0005$. For LAM, the non-clustering algorithm (Median = 0.0499) significantly outperformed the clustering algorithm (Median = 0.0524), $z = 3.920$, $p < .0005$. With the exception of F-score, these results reinforce our findings in Phase 1, where we had used a larger number of participants but fewer topics. The F-score, which represents the trade-off between recall and precision, indicates the difference in recall between the two algorithms is larger than the difference in precision in this second phase.

Table 55 provides the evaluation efficiency for each of our two algorithms as compared with the pooling assessment method used in TREC. In Phase 2, we achieve average F-scores that are 448% and 253% of the TREC pooling method for the non-clustering decision approaches, respectively. As can be inferred from Table 55, assuming the costs for TREC assessors and our methods are equal, the effectiveness of our methods in this deeper study is greater than the pooling method used by TREC.

Table 55: Precision, recall, and F-score of each phase, as compared with the TREC pooling method

Document Collection	# Topics	TREC Assessors			Non-clustering			Clustering		
		Prec	Rec	F	Prec	Rec	F	Prec	Rec	F
Phase 1	3	0.053	1.000	0.099	0.305	0.656	0.416	0.138	0.780	0.235
Phase 2	20	0.049	1.000	0.093	0.324	0.697	0.442	0.145	0.817	0.246

Table 56 presents each phase on the same three topics evaluated by precision, recall, and F-score. This illustrates how the clustering, non-clustering and TREC pooling approaches compare in terms of overall efficiency.

Table 56: Precision, recall, and F-score of each algorithm across News topics, as compared with the TREC pooling method

Collection	News Topic #	TREC Assessors			Non-clustering			Clustering		
		Prec	Rec	F	Prec	Rec	F	Prec	Rec	F
Phase 1	403	0.020	1.000	0.039	0.200	0.952	0.331	0.030	0.952	0.058
	421	0.047	1.000	0.090	0.070	0.241	0.108	0.050	0.639	0.093
	436	0.092	1.000	0.168	0.140	0.211	0.168	0.090	0.606	0.157
Phase 2	403	0.020	1.000	0.039	0.300	0.857	0.444	0.071	0.952	0.133
	421	0.047	1.000	0.090	0.128	0.277	0.175	0.069	0.518	0.122
	436	0.092	1.000	0.168	0.236	0.578	0.335	0.140	0.639	0.230

From Table 56, we observe that the precision from the documents assessed, the precision of both the non-clustering and clustering algorithms is higher when the entire document is considered for assessment (as in Phase 2), as opposed to only using the first 7 sentences of text (as in Phase 1). There is also an increase for both algorithms in recall and F-score between phases, but this increase is relatively smaller than the increase in precision when the entire document is used for assessment.

Similar to what we did for each collection in Section 6.11.3, we rank the 11 clusters used in the clustering approach by their mean weighted score. This is shown in Figure 16. We remove clusters from lowest score to highest score and evaluate precision, recall, and F-score for each collection.

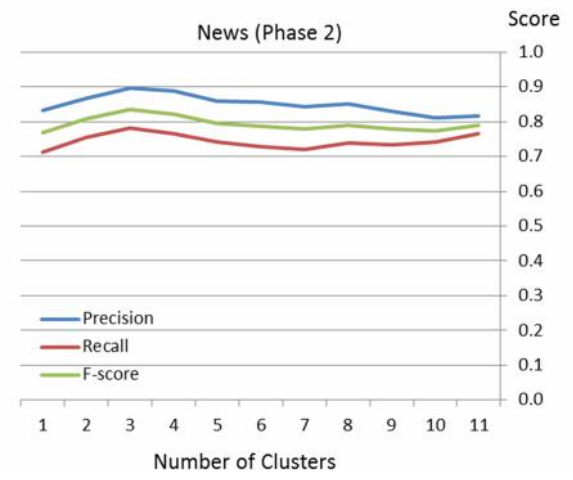

Figure 16: Measures of performance as the number of clusters are increased for the News collection in Phase 2

From Figure 16, in each graph, we observe that when all 11 clusters are evaluated, our F-score (middle line, in green) is not at its peak. The optimal F-score is obtained when we evaluate only the first 3 clusters and ignore the rest, a finding consistent with Phase 1. In Table 57, we examine the mean weighted score and the overall mean for each collection and

find that if we eliminate clusters with a mean weighted score (wtscore) that is below the overall mean for the collection, we can increase our overall performance. Consistent with what was found with the News collection in Phase 1, this threshold would have the crowd evaluate only the 3 highest-ranking clusters in Phase 2 to optimize performance.

Table 57: Ranking of clusters by mean weighted score for Phase 2

Cluster	Mean wtscore	% of Mean wtscore
1	*1571.09*	*1.86*
2	*1021.66*	*1.21*
3	*885.36*	*1.05*
4	799.30	0.94
5	793.45	0.94
6	780.60	0.92
7	766.70	0.91
8	756.83	0.89
9	736.04	0.87
10	719.94	0.85
11	483.45	0.57
Collection Mean	846.77	1.00

6.12 Conclusion

Many relevance assessment campaigns, such as those sponsored by TREC, use a pooling approach to find relevant documents while limiting the costs involved with assessment. In this study, we provide an approach that combines algorithmic and human computation methods to lower costs relative to pooling while still achieving high quality. We find that our methods are able to locate a majority of documents in two types of collections at a fraction of cost of pooling. We first use one of two algorithms to find the documents most likely to be relevant. These documents are then assessed by humans through a crowdsourcing platform. By combining these methods, we are able to have humans and algorithms carry out the tasks they perform best.

We examine two algorithms: a non-clustering algorithm, which uses information gained from the run submissions and a clustering algorithm, which obtains information from the document text in addition to the information used in the non-clustering method. We find that the clustering method improves recall, while the non-clustering method improves precision and LAM. We also describe how the use of a threshold on the clustering algorithm can improve precision and recall. Examining the News collection in depth in our second phase, we also find using the entire document text, as opposed to the first 7 sentences, increases all of our performance measures. Each method examined here provides a cost-effective method to obtain a majority of the relevant documents that would be found by the pooling process while considerably reducing the cost of obtaining them.

CHAPTER 7

QUERY FORMULATION STUDY

Although searching is a core component of any document retrieval system, few user information needs are satisfied by the initial query. In studies of Web searches, which parallel document searches, more than half of all queries are subsequently reformulated by users after results are returned from an initial query (Spink, Jansen, Wolfram, & Saracevic, 2002). Query refinement is often necessary due to the presence of over- or under-specified search terms, inappropriate terms retrieving non-relevant documents, and typos. Thus, query refinement is an important step and a core area of study in IR.

The difficulty with the initial query and query refinement may be due to inadequate guidance; most users receive little, if any, instruction on designing effective queries and also have difficulty identifying useful terms for effective query expansion (Ruthven, 2003). Since users are typically unaware of the depth or the contents of the document collection in advance, they are neither able to measure (or estimate) their own search success nor are they able to compare their own results with those of others searching the same collection. This results in few opportunities for users to improve their search techniques in an objective manner. This in turn, potentially leads to the perpetuation of these same search-related errors on subsequent queries.

7.1 Background and Motivation

Given how important it is to have an effective query for document retrieval it is not surprising that query design, term expansion strategies, methods for reformulating term weights etc., have been explored over the last several decades. There are many studies involving algorithmic methods (such as the classic Rocchio algorithm (Rocchio, 1971) and classifiers (Joachims, 1996)) and many others exploring human intelligence (using expert searchers and librarians, e.g., (Dillon & Song, 2006; McKibbon, *et. al.*, 1990; Turtle, 1994)). At this point it is almost universally acknowledged that in most cases an initial query refined

using a reasonable strategy will yield better results than the initial query. The basis of the refinement may be true or pseudo relevance feedback derived from the documents retrieved by the initial query.

Our goal is to assess the use of human computation through crowdsourcing and GWAPs both for initial query design and for query refinement using feedback. This examination of human computation is not that of the original user or of an expert librarian (an angle well-studied in the literature), but of the largely anonymous individuals. As indicated in (C. G. Harris & P. Srinivasan, 2012), if the methods examined here are found to be effective then we will have the beginnings of a new approach for assisting searchers with query design. This option may be invoked when a query is particularly difficult and the information need has longevity (e.g., in topic detection and tracking (Allan, Papka, & Lavrenko, 1998)) or where some latency in returning results can be tolerated.

We study the value of using largely anonymous people via crowdsourcing for query design; this includes both initial query formulation and query refinement given some relevance feedback. We study this anonymous people approach in game (GWAP) and non-game settings. This allows us to tease out, for example, the effects of offering entertainment on quality and cost. As a contrast we also study query design with a more homogenous and not so anonymous group of individuals; namely students in a campus. Finally we compare performance with an algorithmic baseline. We compare retrieval results obtained using all of these query design methods applied to a common set of topics and by running the resulting queries with the same retrieval algorithms and against the same collection.

This third study examines Step 8 (for the initial query) and Step 11 (for the query refinement with feedback) in our IR model. It involves the formulation of an initial query to meet an information need, and using the results of feedback provided to the user to aid in query reformulation. Although creating initial queries and query reformulation to match information needs has been studied for decades, there have been remarkably few experiments using games or crowdsourcing reported in the literature to date.

7.1.1 Crowd-based approaches

There has been very little empirical research on using crowds to help formulate queries. Integrating the crowd is becoming more commonplace for the difficult searches, perhaps indicating the crowd represents a nice tradeoff between speed, cost, and quality. Bozzon *et. al.* describe a tool called CrowdSearcher, which utilizes the crowd for difficult searches in (Bozzon, *et. al.*, 2012) , but its performance against an established baseline was not examined and thus we are unable to assess their methods against established algorithms. A study by Yan *et. al.* described a mobile search application in (Yan, *et. al.*, 2010); claiming a search precision of 95% but this was not examined empirically. Ageev *et. al.* conducted an experiment to evaluate crowd search techniques in (Ageev, *et. al.*, 2011), but do not compare the crowd's performance with other groups or algorithmic approaches. Several studies, such as (Omar Alonso & Mizzaro, 2012; McKibbon, *et. al.*, 1990) have compared the crowd to experts in document assessment, concluding there is little difference in quality, particularly when multiple assessors are used.

These studies provide the premise that the crowd can be used to search effectively and deliver results with reasonable precision and thus motivate our goal to examine the crowd's effectiveness in initial query formulation. Likewise, there has not been any research in the literature on the ability to refine queries. In our study, we will compare the crowd's ability to provide query terms and phrases against an algorithm approach that has performed well in previous studies, as well as against students. This also motivates this study, as we wish to see if the crowd is as effective, or more effective, than students or established algorithmic approaches.

7.1.2 Game-based approaches

Very few Games With A Purpose (GWAP) have been constructed to address initial query and query reformulation effectiveness. Thumbs-up (Dasdan *et. al.*, 2009) is a GWAP that uses output-agreement mechanism to gather relevance data. This game asks players to

evaluate search terms and attempt to independently determine the most relevant document to a given query. Search War (Edith Law, *et. al.*, 2009) is another game used to obtain data on search relevance and intent for a user-provided query. Players are paired and each given a unique search query and the objective of guessing their opponent's search query first. The design relies on the premise that players will select the least relevant webpage w.r.t. the search query, to provide to their opponent as hints, which implicitly provides a relevance judgment.

Koru (Milne, *et. al.*, 2008), the most similar game to the one we use in our study, allows users to assess their search skills relative to other searchers and evaluate how their own searches might be improved. Like other GWAPs, it is intended to be both fun and to create valuable output on query refinement behavior in a controlled information task. However, it does not make a comparison between different approaches and it is limited to a small document collection from a single source (the New York Times). It also did not evaluate the performance of participants against an established algorithmic baseline. In our study, we will examine the performance of games against the performance of algorithms

7.1.3 Machine based approaches

There have been a number of studies that examine interactive query expansion versus automatic query expansion and reformulation. Interactive query expansion and reformulation can be used as an effective means of improving a search. Efthimiadis (Efthimiadis, 2000) found system-provided terms, on average, when selected, improved retrieval performance. Conversely, Belkin, *et. al.* (Belkin *et. al.*, 2001) found that humans rarely used relevance feedback features and were often puzzled by some machine-suggested terms. Ruthven (Ruthven, 2003) demonstrated that human searchers are less likely than Algorithmic systems to make good expansion and reformulation decisions. Anick (Anick, 2003) found that users made little use of machine-suggested terms to expand and refine their queries, but when they did it improved retrieval performance. Thus, there are mixed performance results from

machine-provided query reformulation and these Algorithmic approaches have not been adequately evaluated against human computation-based methods.

In our study, we apply a strategy used in several TREC submissions. Submitted runs using this query expansion approach consistently outperformed other teams across different IR metrics for the terabyte and ad hoc tracks (Metzler, Strohman, Turtle, & Croft, 2004)

7.2 Contributions

We study the value of using largely anonymous people via crowdsourcing for query design; this includes both initial query formulation and query refinement given some relevance feedback. We study this anonymous people approach in game (GWAP) and non-game settings. This allows us to tease out the effects of offering entertainment on quality and cost. As a contrast we also study query design with a more homogenous and pseudo-anonymous group of individuals; namely students in a campus. Finally we compare performance with an algorithmic baseline. We compare retrieval results obtained using all of these query design methods applied to a common set of topics and by running the resulting queries with the same retrieval algorithms and against the same collection.

7.3 Hypotheses

In this study, we wish to examine the following eight hypotheses. For initial query and for query reformulation with feedback, we examine precision across the top 10 retrieved documents (p@10), average precision, recall and F-score.

7.3.1 Initial Query Precision@10

$H_{0\text{-}1P}$: there are no main or interaction differences in p@10 in the initial query due to the type of interface, type of participant, or type of collection used.

$H_{A\text{-}1P}$: there are main or interaction differences in p@10 in the initial query due to the type of interface, type of participant, or type of collection used.

7.3.2 Initial Query Average Precision

$H_{0\text{-}1A}$: there are no main or interaction differences in mean precision in the initial query due to the type of interface, type of participant, or type of collection used.

$H_{A\text{-}1A}$: there are main or interaction differences in mean precision in the initial query due to the type of interface, type of participant, or type of collection used.

7.3.3 Initial Query Recall

$H_{0\text{-}1R}$: there are no main or interaction differences in mean recall in the initial query due to the type of interface, type of participant, or type of collection used.

$H_{A\text{-}1R}$: there are main or interaction differences in mean recall in the initial query due to the type of interface, type of participant, or type of collection used.

7.3.4 Initial Query F-score

$H_{0\text{-}1F}$: there are no main or interaction differences in mean F- in the initial query due to the type of interface, type of participant, or type of collection used.

$H_{A\text{-}1F}$: there are main or interaction differences in mean F-score in the initial query due to the type of interface, type of participant, or type of collection used.

7.3.5. Query Refinement with Feedback Precision @10

$H_{0\text{-}2P}$: there are no main or interaction differences in mean precision in query refinement with feedback due to the type of interface, type of participant, or type of collection used.

$H_{A\text{-}2P}$: there are main or interaction differences in mean precision in query refinement with feedback due to the type of interface, type of participant, or type of collection used.

7.3.6. Query Refinement with Feedback Average Precision

$H_{0\text{-}2A}$: there are no main or interaction differences in mean precision in query refinement with feedback due to the type of interface, type of participant, or type of collection used.

$H_{A\text{-}2A}$: there are main or interaction differences in mean precision in query refinement with feedback due to the type of interface, type of participant, or type of collection used.

7.3.7. Query Refinement with Feedback Recall

$H_{0\text{-}2R}$: there are no main or interaction differences in mean recall in query refinement with feedback due to the type of interface, type of participant, or type of collection used.

$H_{A\text{-}2R}$: there are main or interaction differences in mean recall in query refinement with feedback due to the type of interface, type of participant, or type of collection used.

7.3.8 Query Refinement with Feedback F-score

$H_{0\text{-}2F}$: there are no main or interaction differences in mean F-score in query refinement with feedback due to the type of interface, type of participant, or type of collection used.

$H_{A\text{-}2F}$: there are main or interaction differences in mean F-score in query refinement with feedback due to the type of interface, type of participant, or type of collection used.

7.4 Datasets and Topics

7.4.1 Document Collections

General Document Collection (News): Data collection used for the TREC 2012 Crowd TRAT subtask (disk 4 and 5) – a general News collection. We use the 18,260 test documents for the TREC-2012 Crowd TRAT subtask. The number of relevant documents per topic ranged from 7 (for topic 380) to 361 (for topic 354), with an average of 87.9 relevant documents per topic.

Specialized Document Collection (OHSUMED): TREC-9 filtering track dataset – a specific collection of OHSUMED (MESH) abstracts, titles, and MeSH (Medical Subject Heading) terms (Hersh, *et. al.*, 1994). We use the 293,550 test documents from 1988-1991. The number of relevant documents per topic ranged from 12 (for topic 4) to 172 (for topic 30), with an average of 68.8 relevant documents per topic.

7.4.2 Topics

General Document Collection (News): we randomly selected 20 topics used in the TREC-7 ad hoc task. Outdated topics were discarded. The 20 topics chosen were: 351, 354, 355, 358, 359, 363, 364, 369, 374, 375, 379, 380, 388, 389, 390, 393, 395, 396, 399, and 400. These were presented to each user in the same order.

Specialized Document Collection (OHSUMED): TREC-9 filtering track test dataset – the OHSU-created topics 1-43. We randomly selected 20 topics from the 43 OHSUMED topics used in the TREC-9 filtering task. The 20 topics chosen were: 3, 4, 6, 9, 11, 12, 13, 15, 16, 18, 20, 21, 22, 24, 28, 30, 33, 35, 36, and 41.

7.5 Gold standard

For the News collection, we obtain our gold standard judgments using a subset of topics and the judged assessments from the TREC-7 and TREC-8 ad hoc tasks. The gold standard judgments are the binary relevance assessments used in this task.

For the experiment against a specialized data collection, we obtain our gold standard using the relevance judgments provided by OHSUMED assessors on the TREC-9 filtering track test dataset as our gold standard. This is the same gold standard used with the relevance assessment study described in Chapter 6.

7.6 Modes of Interaction

We use two different modes of interaction for this task: a game interface and a non-game interface. Each of these interfaces has two phases: the initial query and the query refinement with feedback.

For the non-game interface, we use a tool called “Seek-o-rama”. For each topic, Seek-o-rama presents the users with the title, an information need (a slight modification of the description phrased to indicate an information need), and the narrative, which provides additional constraints for the information need for each topic. The user inputs the query

terms and operators. The user's query is stored and run using Indri in a 2 hour window between phases. A screenshot is shown in Appendix H.

In the query reformulation phase, the user is given precision and recall information to provide feedback on their performance, along with their highest ranking relevant document and highest-ranking non-relevant document (which are clearly marked). The user can use this information to refine their query terms and improve their score.

For the game format, we use a tool called "Seekgame". As with Seek-o-rama, the Seekgame presented users with the title, information need and the narrative. The user inputs the query terms and operators to match this information need. The query terms are stemmed, stop listed, and matched against a database of terms parsed from the headline and document text of a randomly selected relevant document for each topic. Matching more terms stored in the database indicated a better match. The user's score is provided immediately, along with the highest sore achieved for that topic as a benchmark. The user is timed for each topic and given a point bonus for providing matching terms quickly. Screenshots are shown in Figures G-1 to G-4 in Appendix G.

In the second, query refinement phase, the user is presented the highest relevant and non-relevant document obtained using the baseline algorithmic approach. The user is then able to refine their query and improve their score using this information. The user is timed and provided with instant scoring as was done in the first phase.

At the conclusion of each phase of Seekgame, users are provided with a leaderboard and awarded badges and stars for achieving a high score during each of the two phases.

7.6.1 Seek-o-rama (Data Collection Web Interface)

To examine queries issued through standard browser interface, we invited participants to use Seek-o-rama, a PHP-based data collection interface.

7.6.1.1 Initial Query Formulation

Users were provided with the title, the description, and the narrative for each of the 20 topics, one at a time. Participants were given a large text box to input their query, with a pop-up help screen available to them throughout the task.

7.6.1.2 Query Refinement

The user's original search terms were pre-loaded in the input text boxes for each topic, allowing easy modification to their original query. Also, in the second round, users were provided with the highest-ranked relevant and non-relevant document from the collection to aid them in their query refinement.

7.6.2 Seekgame (Game Interface)

Some users invited to participate in this exercise were randomly selected to use Seekgame, a PHP-based game, instead of the Seek-o-rama interface.

7.6.2.1 Initial Query Formulation

Partiicpants selected to use Seekgame were given a different URL, and were presented with the same initial screen outlining the game's objectives, instructions on term and operator rules as the Seek-o-rama interface participants.

The game instructions also had the following additions. First, there was a time-based constraint that required search terms to be entered within 30 seconds. Second, scoring was provided instantly (explained soon). Third, participants had musical sound effects to enhance the interface's game-like feel. Last, a leader-board and badges, or icons, were awarded for superior game performance.

7.6.2.2 Query Refinement.

Unlike Seek-o-rama, the Seekgame did not provide users with precision and recall information from their initial round as they began their second round. This was because the calculation of this information was not integrated into the game interface and would take

away from the feeling of engagement. Instead once a user entered a set of terms for a topic, these terms were parsed to remove stop words, stemmed, and compared against a weighted list of stemmed terms obtained from documents judged relevant for that topic. The list of stop words used appears in Appendix E.

A pop-up screen provided scoring and bonus information to each player after they submitted their query. This score was immediately calculated and issued to the user, along with a time-based bonus for completing the search quickly. Once a user completed the first round, they could begin the query refinement round without delay. Users were instructed to refine their initial query based on their score and a relevant and non-relevant document provided to them to aid their refinement, subject to the same 30-second time restriction.

Stars were awarded to users who scored above a certain threshold. Virtual badges were given to users having the highest overall score, and a leaderboard was shown to the users, providing the option for top scorers to add their names for "bragging rights".

7.7 Scoring

7.7.1 Initial Query (Round 1)

For OHSUMED collection, they are given the following instruction in Round 1. Actual screenshots of the instructions appear in Appendix G.

> OBJECTIVE
>
> The objective of Seek-o-rama is to help us find relevant documents for topics. You will help us by entering phrases that can be used to search for relevant documents for the topics given to you. There are two rounds. In each round, you will do the same thing: enter for each topic one or more phrases, separated by semicolons (;). For example, if you want to enter both the phrases 'high blood pressure' and 'hypertension' these may be entered as:
>
> high blood pressure; hypertension
>
> Entering your phrases in uppercase or lowercase does not affect your score.

> You will be given 20 topics.
>
> ROUNDS:
>
> You are shown each topic and asked for search phrases. At the conclusion of Round 1, we will use your phrases to construct queries. Your score for each topic will be based on the quality of the returned documents. During Round 2, you can see your score from Round 1. You may then make modifications to improve your score. You will also be given the headline/title from a relevant document for each topic. Clicking on the headline/title of these documents links to the document abstract, which will appear in a pop-up window. This document may help you enter better phrases in Round 2.
>
> SCORING:
>
> The score for your query, in terms of finding all relevant and only relevant terms, is a percentage between 0 and 100 - a higher percentage is better. At the end of Round 1, your score will be displayed for each topic. Your objective in Round 2 will be to improve upon your Round 1 scores.
>
> NOTE ABOUT POP-UPS:
>
> Relevant documents will appear in a pop-up screen in Round 2. On some browsers, if the pop-up is not closed properly, the next time the document title link is clicked, the pop-up containing the text will appear behind the current browser window. If the pop-up does not appear, please see if it is behind your current browser window.

The instructions for the News collection are close to identical, but use a different example in the objective and indicate it uses the first 7 sentences from the text, not the full abstract). The example used is as follows:

> For example, if you want to enter both the phrases ‘carbon monoxide’ and ‘global warming’ enter them as:
>
> carbon monoxide; global warming

7.7.2 Query Refinement Based on Feedback (Round 2)

In the second round, users are given the following instruction for the OHSUMED collection, and a similar one for the News collection:

> The following are the phrases entered in your initial Round 1 search. In this round, we again provide the information need you were shown before and also include the title and the full abstract for a document that is relevant to each information need.
> You may modify the phrases you entered earlier, as you feel appropriate, in order to improve your search. You may also add new phrases or modify or even remove phrases you provided

earlier. Once again, please separate phrases by a semicolon (;). For example, a search that includes both the phrases 'high blood pressure' and 'hypertension' may be entered as:

High blood pressure; hypertension

7.7.3 Game Approach

The game approach uses the same instructions for each collection as the non-game version, but provides a different scoring mechanism. The information used in the experiment will be parsed the same, weighted the same, and evaluated in the same manner through Indri as the non-game version. The difference will be the scoring information provided to the user for instant feedback. We align this score with the actual task performance as closely as possible, but we are limited by the fact that we must do it in real-time.

In addition to the randomly-selected relevant document for each topic used in our experiment, we randomly-select a second relevant document for each topic, called the *game scoring document*, which is used in the game scoring. For each topic, we store the single non-stoplisted words obtained from the game scoring document.

Scores are determined by the frequency of occurrence in the title/headline and abstract for OHSUMED (or title/headline and first 7 sentences of text for News) in the game scoring document. If the phrase entered by participant contains one or more of these single words, the participant's game score will be increased according to the number of times each word appears in the document. For example, if a participant supplied 3 non-stoplisted words, each of which appeared 4 times in the game scoring document, that participant would receive a score of 12 points for that document. The score from the game and the overall performance metrics correlated well - a post-hoc examination found the correlation between the score participants achieved in the game and their precision and recall scores was 0.641 and 0.596 respectively ($p < 0.001$, $df = 1438$).

To make the game more challenging, we also provide game participants with a scoring bonus for entering phrases quickly. This time-based bonus is added to the word-matching score. For example, if the participant earned bonus of 12 points for entering

words correctly for a topic, had 10 seconds remaining on the countdown timer, and the points-per-second factor is 0.05, the user would earn a bonus of 6 points.

Base score = 12 points

Time-based bonus = seconds remaining x points-per-second factor x base score

= 10 x 0.05 x 12

= 6 points

Total score earned = 12 + 6 = 18 points

This bonus is a percentage of the word-based score earned, ensuring the participants maintain a focus on providing meaningful phrases. For the actual query that is evaluated in our study, we evaluate phrases as we do with the non-game version. For example, in the game interface, if the user entered the following phrases to match the information need

> tennis; Nadal; grand slam; Wimbledon

If one of the three highest-ranked documents had the following title:

> Nadal beats Federer to win his Second Grand Slam Title

If the text in the game-scoring document is:

> Nadal beat Federer in three games to win his second men's Grand Slam title at Wimbledon Friday. The tennis tournament was highly anticipated rematch between the two. Federer and Nadal had also met earlier this year in the French Open.

Our terms for success (and points used in success scoring) would be:

- Nadal 3
- Beat 2
- Federer 3
- Win 2
- Grand 2
- Slam 2
- Title 2
- Game 1
- Wimbeldon 1
- Friday 1
- Tennis
- Tournament 1

- Highly 1
- Anticipated 1
- Rematch 1
- Met 1
- Earlier 1
- Year 1
- French 1
- Open 1

Using these terms and scores will provide the participant a word matching score of 1 + 3 + 2 + 2 + 1 = 9. If the game participant hit the 'submit' button with 6 seconds remaining, and the point-per-second-factor was 0.05 points, they would earn a bonus of (9 x 6 x 0.05) = 2.7 points, for a total of 11.7 points. From this example, the phrases for Indri evaluation in the experiment would be:

Tennis
Nadal
Grand slam
Wimbledon

These phrases would be evaluated against the same randomly-selected document as is used in the non-game version.

In Round 2, we provide the participant with their Round 1 query, provide them with the same randomly-selected relevant document as provided to non-game participants and allow them to refine their query. We explain scoring in the game version as follows:

> SCORING:
>
> Points are awarded in each round for providing phrases with words which match those in relevant documents. A scoring bonus is given for the number of seconds left on the countdown timer when the submit button is pressed. The scoring bonus is given as a percentage of the points earned for matching words that query. Therefore, meaningful phrases must be entered to score bonus points. In Round 2, you are provided with your Round 1 query and your objective is to modify your query to improve upon your Round 1 score.

7.8 Participants

We use participants from the crowd and students in this task. Students are undergraduate volunteers solicited in an undergraduate economics course. Both participants are randomly assigned to using the standard web interface (Seek-o-rama) or the game interface (Seekgame). Crowdsourcing participants are hired using Amazon Mechanical Turk and compensated $0.20 for participation and asked to provide twenty queries based on an

information need. They were given another $0.20 to provide refinements based on feedback to their original queries based on the initial information need.

Each participant was randomly assigned to a mode of interaction (either the game interface or the non-game interface) and to a collection (either the News collection or the OHSUMED collection). We had 18 participants for each combination of collection, interface, and participant type.

In addition to recording their IP address for geo-location purposes, we also required participants to provide some information prior to beginning the task. Table 57 provides the responses to each survey categories asked of participants and the percentages responding to each choice.

Table 58: Demographic information obtained from participants. Percentage indicating each choice is given in parentheses.

Category	**Participant Response**
Region (as determined by IP address)	North America (57.2%), Europe (13.2%), Africa/Middle East (1.9%), South Asia (20.9%), East Asia (4.4%), South America (0.4%), Australia/NZ/Oceana (2.0%)
Age	<18 (14.6%), 19-25 (57.0%), 26-35 (25.6%), 36-45 (2.8%), 46+ (0%)
Gender	Male (47.9%), Female (52.1%)
English Ability	Poor (0.2%), Moderate (14.2%), Good (29.6%), Fluent (56.0%)
Education	No baccalaureate (64.3%), Completed baccalaureate (35.7%)
Current Student?	Yes,Full-time (61.1%) , Yes,Part time (30.2%), No (8.7%)
Chemistry course in last 5 years?	Yes (48.8%), No (51.2%)

7.9 Procedure

7.9.1 Initial Query (Phase 1)

For each topic, the title and information need are broken into phrases based broken on punctuation and on stop words, such as conjunctions and prepositions. This approach is the same for both the News and OHSUMED collections.

Using Topic 32 as an example, the following is provided to the user:

Title:
42 year old black male with hypertension

Information Need:
Find documents that describe the utility of beta blockers and blacks with hypertension

Using the 635-word stop word list[34], we obtain the following phrases:

- year
- black male
- hypertension
- utility
- beta blockers
- blacks
- hypertension (a duplicate, which is removed)

We determine weights for each component of the query by replicating the query expansion weighting examples provided in (Metzler, Strohman, Turtle, & Croft, 2004). These examples follow the strategy used by the UMass team in several TREC submissions as closely as possible. The UMass team's submitted runs using this query expansion approach consistently outperformed other teams and the baseline across different IR metrics for the 2004-2006 Terabyte track and ad hoc tracks (Metzler, Strohman, & Croft, 2006). The UMass team also reports the use of this query expansion strategy and weights provide a significant improvement over their baseline ad hoc runs in TREC (Metzler, *et. al.*, 2004). The weights we use for each section of our documents are consistent with their approach and are provided in Table 59.

[34] http://www.webconfs.com/stop-words.php (the list is also provided in Appendix E)

Table 59: Weights assigned to each component of the Indri query.

Initial Query	News	OHSUMED
Ordered phrases in information need and description/title	1.5	1.5
Ordered bigrams in information need and description/title	0.1	0.1
Unordered bigrams in information need and description/title	0.3	0.3
Revised Query	**News**	**OHSUMED**
Ordered phrases in the document title/heading and document abstract	N/A	1.5
Ordered bigrams in the document title/heading and document abstract	N/A	0.1
Unordered bigrams in the document title/heading and document abstract	N/A	0.3
Ordered phrases in the document title/heading and in the first 7 sentences of document text	1.5	N/A
Ordered bigrams in the document title/heading and in the first 7 sentences of document text	0.1	N/A
Unordered bigrams in the document title/heading and in the first 7 sentences of document text	0.3	N/A

The UMass query expansion strategy is as follows. First, each phrase is evaluated in its entirety. Second, the phrase is broken into bi-grams: (a) ordered combinations of bi-grams, (b) unordered combinations of bi-grams appearing within an 8-word window in each document. Last, we examine an unordered set of all words in the phrase appearing within a 12-word window in each document.

The document provided by UMass describes an example. Given the query "Prostate cancer treatments" (topic 710) the UMass system generates the following query:

```
#weight(
1.5 #combine( prostate cancer treatments )
0.1 #combine(
                #1(cancer treatments)
                #1(prostate cancer)
                #1(prostate cancer treatments))
0.3 #combine( #uw8(cancer treatments)
                #uw8(prostate treatments)
                #uw8(prostate cancer)
                #uw12(prostate cancer treatments))
)
```

To generalize this approach, given a phrase with words A B C:

```
#weight (
X       #combine (A B C)
X/15    #combine (#1 (A B) #1 (B C) #1(A B C)
X/5     #combine (#uw8(A C) #uw8(A B) #uw8(B C)
        #uw12( A B C))
)
```

where X is the weight of the entire ordered phrase. We set X at 1.5 - the same as the UMass team used in their successful runs. Note that all weights are relative. In Indri's INQUERY language, #1 indicates an ordered phrase, #uwK indicates an unordered window of size K (for additional information about INQUERY query constructs, we refer the reader to Strohman, Metzler, Turtle, & Croft, 2005). This is counter-intuitive as typically a higher weight is given to ordered bigrams than unordered windows.

Thus, in the previously-provided example, OHSUMED Topic 32, we would provide the following Indri query term:

```
#weight (
   1.5 #combine (year)

   1.5 #combine (black male)
   0.1 #combine (#1(black male))
   0.3 #combine (#uw8(black male))

   1.5 #combine (hypertension)

   1.5 #combine (utility)

   1.5 #combine (beta blockers)
   0.1 #combine (#1(beta blockers))
   0.3 #combine (#uw8(beta blockers))

1.5 #combine (blacks)
  )
```

7.9.2 Feedback-based Query Refinement (Phase 2)

We calculate the number in sentences contained in the OHSUMED abstracts and find they contain, on average, 6.97 sentences. To make a comparably-sized portion for review from the News collection, we use the first 7 sentences of the text.

We randomly select a relevant document for each topic from the gold standard. For example, OHSU document 88279324 is a randomly-selected relevant document selected from the gold standard for OHSUMED topic 32. For a given topic, the same relevant document is provided to both the algorithm and shown to each participant. The document shown to participants and the algorithm is exactly the same.

<DOC>
<DOCNO>88279324</DOCNO>
<HEADLINE>
Secondary prevention in elderly survivors of heart attacks.
</HEADLINE>
<TEXT>
More than 200,000 elderly patients survive myocardial infarctions each year. Thus, the achievement of even minimal decreases in reinfarction and mortality rates will benefit large numbers of patients. Secondary prevention strategies include smoking cessation; the control of hyperlipidemia, obesity and diabetes; the management of hypertension and stress; exercise; the use of drugs such as beta blockers and aspirin, and increased attention to general health.
</TEXT>
</DOC>

. For the OHSUMED collection, we evaluate the headline/title and the document abstract; for the News collection, we use the headline/title and the first 7 sentences of the text. We extract phrases and extract stop words using the same algorithm described in Appendix C in the same manner as we did in Phase 1.

Phrases are extracted, broken on stop words and on punctuation. Numbers in the text are treated as stop words. We obtain the following phrases from document 88279324:

- secondary prevention
- elderly survivors
- heart attacks
- elderly patients survive myocardial infarctions
- year
- achievement
- minimal decreases
- reinfarction
- mortality rates
- patients
- secondary prevention strategies
- smoking cessation
- control
- hyperlipidemia obesity
- diabetes
- management
- hypertension
- stress
- exercise
- drugs
- beta blockers
- aspirin
- increased attention
- general health

We append the above to our Round 1 query.

We use the weights from Table 55. Using these weights, our query for Topic 32 becomes:

```
#weight (
   1.5 #combine (secondary prevention)
   0.1 #combine (#1(secondary prevention))
   0.3 #combine (#uw8(secondary prevention))

   1.5 #combine (elderly survivors)
   0.1 #combine (#1(elderly survivors))
   0.3 #combine (#uw8(elderly survivors))

   1.5 #combine (heart attacks)
   0.1 #combine (#1(heart attacks))
   0.3 #combine (#uw8(heart attacks))

   1.5 #combine (
                  elderly patients survive myocardial infarctions)
   0.1 #combine (
         #1(elderly patients)
         #1(patients survive)
```

```
        #1(survive myocardial)
        #1(myocardial infarctions))
0.3 #combine (
        #uw8(elderly patients)
        #uw8(elderly survive)
        #uw8(elderly myocardial)
        #uw8(elderly infarctions)
        #uw8(patients survive)
        #uw8(patients myocardial)
        #uw8(patients infarctions)
        #uw8(survive myocardial)
        #uw8(survive infarctions)
        #uw8(myocardial infarctions)
        #uw12(elderly patients survive myocardial infarctions))

1.5 #combine (year)

1.5 #combine (achievement)

1.5 #combine (minimal decreases)
0.1 #combine (#1(minimal decreases))
0.3 #combine (#uw8(minimal decreases))

1.5 #combine (reinfarction)

1.5 #combine (mortality rates)
0.1 #combine (#1(mortality rates))
0.3 #combine (#uw8(mortality rates))

1.5 #combine (patients)

1.5 #combine (secondary prevention strategies)
0.1 #combine (
        #1 (secondary prevention)
        #1 (prevention strategies))
0.3 #combine (
        #uw8(secondary prevention)
        #uw8(secondary strategies)
        #uw8(prevention strategies)
        #uw12(secondary prevention strategies))

1.5 #combine (smoking cessation)
0.1 #combine (#1(smoking cessation))
0.3 #combine (#uw8(smoking cessation))

1.5 #combine (control)

1.5 #combine (hyperlipidemia obesity)
0.1 #combine (#1(hyperlipidemia obesity))
0.3 #combine (#uw8(hyperlipidemia obesity))
```

```
    1.5 #combine (diabetes)

    1.5 #combine (management)

    1.5 #combine (hypertension)

    1.5 #combine (stress)

    1.5 #combine (exercise)

    1.5 #combine (drugs)

    1.5 #combine (beta blockers)
    0.1 #combine (#1(beta blockers))
    0.3 #combine (#uw8(beta blockers))

    1.5 #combine (aspirin)

    1.5 #combine (increased attention)
    0.1 #combine (#1(increased attention))
    0.3 #combine (#uw8(increased attention))

    1.5 #combine (general health)
    0.1 #combine (#1(general health))
    0.3 #combine (#uw8(general health))
)
```

We obtain our machine baseline in the following way. For each topic, we are given a title, description, and narrative. We transform the question in the description into an information need. We then take the title and description, applied stemming, and entered the stop listed terms into Indri, (Strohman, *et. al.*, 2005), the Lemur toolkit for language modeling and information retrieval. This became our baseline initial query.

Using the ranked list returned by Indri, we selected the highest-ranked document from the results of the initial query. We added the terms contained within the headline and byline (subheading) of the retrieved document as additional inputs to the query, applied the stemming and stop word list to the added terms. This became our baseline refined query.

7.10 Results

7.10.1 Initial Query (Phase 1)

Table 60 provides the means and standard deviations for the Initial Query across our four dependent variables

Table 60: Initial query means and standard deviations for dependent variables by interface, collection and participant type (n = 144)

	Initial Query								
	Precision@10		Avg. Precision		Recall		F-score		
Condition	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**N**
Interface									
Game	0.245	0.105	0.223	0.081	0.257	0.084	0.235	0.075	72
Non-game	0.223	0.100	0.214	0.077	0.240	0.085	0.220	0.074	72
Collection type									
News	0.276	0.090	0.231	0.077	0.272	0.077	0.240	0.063	72
OHSUMED	0.192	0.098	0.206	0.081	0.224	0.086	0.215	0.083	72
Participant type									
Crowd	0235	0.110	0.210	0.079	0.250	0.091	0.225	0.080	72
Student	0233	0.095	0.226	0.080	0.247	0.079	0.229	0.069	72
Interface × Collection									
Game, News	0.292	0.086	0.235	0.080	0283	0.073	0.249	0.061	36
Game, OHSUMED	0.199	0.102	0.211	0.082	0.261	0.080	0.221	0.085	36
Non-game, News	0.260	0.092	0.227	0.074	0.231	0.089	0.231	0.064	36
Non-game, OHSUMED	0.186	0.095	0.206	0.081	0.218	0.085	0.209	0.082	36
Interface × Participant									
Game, Crowd	0.242	0.109	0.218	0.077	0.259	0.089	0.235	0.080	36
Game, Student	0.249	0.102	0228	0.087	0.255	0.081	0.235	0.071	36
Non-game, Crowd	0.228	0.113	0.203	0.081	0.240	0.093	0.215	0.080	36
Non-game, Student	0.218	0.086	0.225	0.074	0.238	0.076	0.224	0.068	36
Collection × Participant									
News, Student	0.289	0.087	0.234	0.083	0.256	0.078	0.230	0.061	36
News, Crowd	0.263	0.092	0.228	0.071	0.289	0.072	0.249	0.065	36
OHSUMED, Student	0.203	0.095	0.219	0.077	0.238	0.079	0.228	0.078	36
OHSUMED, Crowd	0.181	0.101	0.193	0.083	0.211	0.091	0.201	0.087	36

Table 60 Continued

	Initial Query								
	Precision@10		Avg. Precision		Recall		F-score		
Condition	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**N**
Interface × Collection × Participant									
Game, News, Student	0.301	0.073	0.244	0.091	0.280	0.071	0.248	0.063	18
Game, News, Crowd	0.284	0.100	0.225	0.069	0.286	0.076	0.250	0.068	18
Game, OHSUMED, Student	0.198	0.103	0.212	0.082	0.230	0.085	0.221	0.083	18
Game, OHSUMED, Crowd	0.199	0.104	0.211	0.086	0.232	0.094	0.221	0.090	18
Non-game, News, Student	0.225	0.085	0.223	0.076	0.231	0.079	0.212	0.061	18
Non-game, News, Crowd	0.294	0.087	0. 231	0.074	0.292	0.070	0.249	0.063	18
Non-game, OHSUMED, Student	0.210	0.089	0.226	0.073	0246	0.075	0.235	0.074	18
Non-game, OHSUMED, Crowd	0.163	0.098	0.174	0.079	0.190	0.086	0.182	0.082	18

7.10.1.1 Initial Query P@10

Table 61: ANOVA results for initial query, P@10

	Precision@10				
Effect	**SS**	**df**	**Mean Square**	**F**	**p-value**
Interface	0.018	1	0.018	12.195	0.001
Participant	0.000	1	0.000	0.013	0.911
Collection	0.251	1	0.251	29.154	0.000
Interface × Collection	0.004	1	0.004	0.447	0.505
Interface × Participant	0.003	1	0.003	0.356	0.552
Collection × Participant	0.021	1	0.021	2.454	0.120
Interface × Collection × Participant	0.040	1	0.040	4.661	0.033
Error	1.172	136	0.009		

Table 61 provides the ANOVA results for precision across the top 10 retrieved documents (p@10) for the Initial Query task. The three-way ANOVA results for p@10 of Initial Query were statistically significant ($p < 0.05$) for collection, with News having a

higher p@10 than OHSUMED, F(1,136)=29.154, p<0.001, for interface, with the game interface having a higher p@10 than the non-game interface, F(1,136)=12.195, p=0.001, and for the interface x participant type x collection interaction, F(1,136)=4.661, p=0.033. This three-way interaction effect is graphically depicted in Figure 17.

Simple effect follow-up tests for the three-way interaction revealed that for students participants evaluating the News collection, the game interface had a higher p@10 than the non-game interface (M_G=0.301, M_{NG}= 0.226; M_{Diff} = 0.095; t(136)=2.380; p=0.019; d=0.79) .

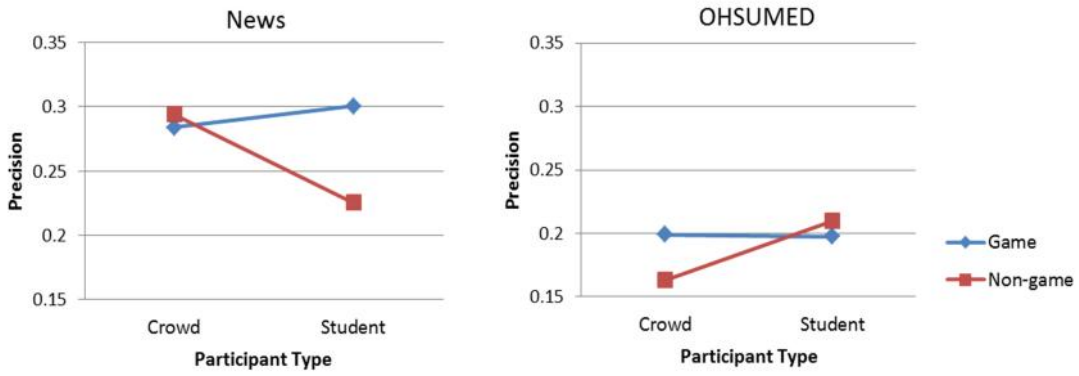

Figure 17: Interaction effects for precision@10 in initial query for interface (lines) and participant type (x-axis) for the News collection (left) and OHSUMED collection (right)

7.10.1.2 Initial Query AveP

Table 62 provides the ANOVA results for Average Precision (AveP) for the Initial Query. The three-way ANOVA results for Initial Query AveP were not statistically significant ($p < 0.05$) for any of the main effects or any two-way or three-way interaction effects between our factors.

Table 62: ANOVA results for initial query average precision

Effect	Average Precision				
	SS	Df	Mean Square	F	p-value
Interface	0.003	1	0.003	0.522	0.471
Participant	0.009	1	0.009	1.501	0.223
Collection	0.022	1	0.022	3.606	0.060
Interface × Collection	0.000	1	0.000	0.016	0.899
Interface × Participant	0.001	1	0.001	0.199	0.657
Collection × Participant	0.004	1	0.004	0.624	0.431
Interface × Collection × Participant	0.013	1	0.013	2.142	0.146
Error	0.847	136	0.006		

7.10.1.3 Initial Query Recall

Table 63: ANOVA results for initial query recall

Effect	Recall				
	SS	df	Mean Square	F	p-value
Interface	0.011	1	0.011	1.716	0.192
Participant	0.000	1	0.000	0.050	0.824
Collection	0.083	1	0.083	12.965	0.000
Interface × Collection	0.001	1	0.001	0.106	0.746
Interface × Participant	0.000	1	0.000	0.001	0.970
Collection × Participant	0.033	1	0.033	5.172	0.025
Interface × Collection × Participant	0.029	1	0.029	4.450	0.037
Error	0.871	136	0.006		

Table 63 provides the ANOVA results for recall for the Initial Query task. The three-way ANOVA results for recall of Initial Query were statistically significant ($p < 0.05$) for the main effect of collection, with News higher than OHSUMED, $F(1,136)=12.965$, $p<0.001$, the

interaction between the collection and participant, F(1,136)=5.172, p=0.025, and three-way interaction interface x participant type x collection interaction, F(1,136)=4.450, p=0.037. This three way interaction, with interface as the primary factor, is presented in Figure 18.

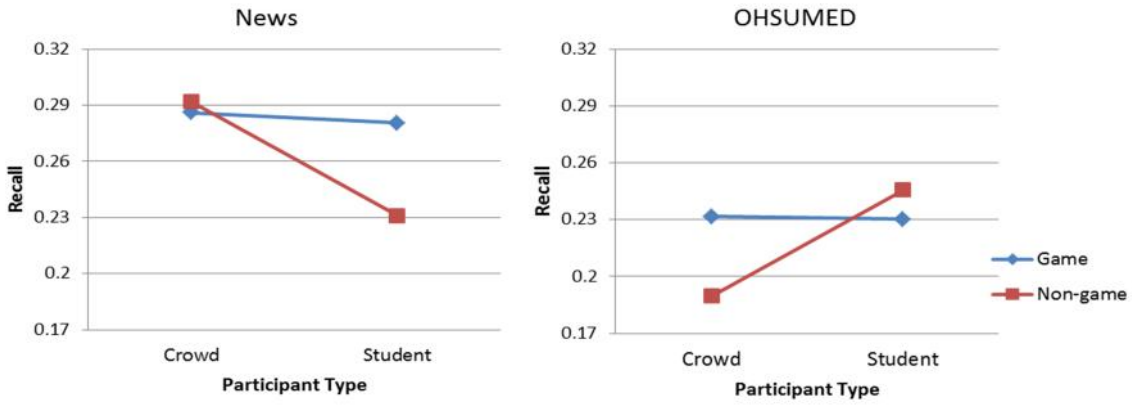

Figure 18: Interaction effects for recall in initial query for interface (lines) and participant type (x-axis) for the news collection (left) and OHSUMED collection (right)

No simple effects were found using interface type as our primary factor, so we evaluate collection type as our primary factor. This three-way interaction effect is graphically depicted in Figure 19.

Simple effect follow-up tests for the three-way interaction found that using the non-game interface with the News collection, crowd participants obtained higher recall than students (M_C=0.292, M_S= 0.231; M_{Diff} = 0.061; t(136)=2.360; p=0.020; d=0.787), whereas using the OHSUMED collection students obtained higher recall than the crowd (M_S=0.246, M_C= 0.190; M_{Diff} = 0.056; t(136)=2.169; p=0.032; d=0.723).

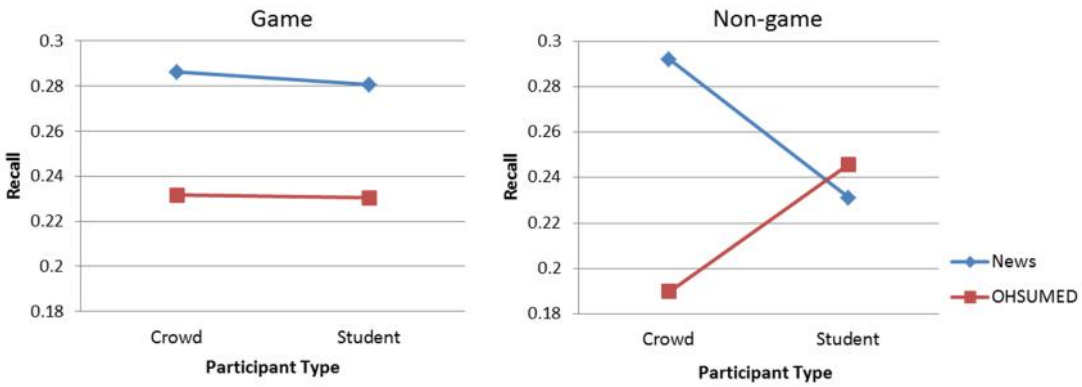

Figure 19: Interaction effects for recall in initial query for collection (lines) and participant type (x-axis) for the game interface (left) and the non-game interface (right)

7.10.1.4 Initial Query F-score

Table 64: ANOVA results for initial query, F-score

Effect	F-score				
	SS	df	Mean Square	F	p-value
Interface	0.008	1	0.008	1.573	0.212
Participant	0.001	1	0.001	0.105	0.746
Collection	0.023	1	0.023	4.259	0.041
Interface × Collection	0.000	1	0.000	0.068	0.795
Interface × Participant	0.001	1	0.001	0.135	0.714
Collection × Participant	0.019	1	0.019	3.546	0.062
Interface × Collection × Participant	0.018	1	0.018	3.357	0.069
Error	0.726	136	0.005		

Table 63 provides the ANOVA results for F-score for the Initial Query task. The three-way ANOVA results for F-score of Initial Query were statistically significant ($p < 0.05$) for the collection, with F-score higher for the News collection than the OHSUMED

collection, F(1,136)=4.259, p=0.041. There were no significant two-way or three-way interactions for Initial Query F-score.

7.10.2 Query Refinement with Feedback (Phase 2)

Table 65 provides the means and standard deviations across our four dependent variables for the Query Refinement with Feedback task.

Table 65: Query refinement with feedback means and standard deviations for dependent variables by interface, collection and participant type (n = 144)

	Query Refinement with Feedback								
	Precision@10		**Avg. Precision**		**Recall**		**F-score**		
Condition	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**N**
Interface									
Game	0.248	0.102	0.235	0.084	0.276	0.091	0.250	0.080	72
Non-game	0.230	0.100	0.224	0.080	0.255	0.092	0.232	0.078	72
Collection type									
News	0.278	0.088	0.245	0.078	0.295	0.083	0.257	0.068	72
OHSUMED	0.200	0.100	0.214	0.082	0.237	0.091	0.225	0.086	72
Participant type									
Crowd	0.244	0.110	0.223	0.084	0.266	0.098	0.238	0.085	72
Student	0.234	0.092	0.237	0.079	0.266	0.084	0.243	0.073	72
Interface × Collection									
Game, News	0.293	0.083	0.249	0.080	0.307	0.079	0.267	0.066	36
Game, OHSUMED	0.204	0.101	0.221	0.086	0.245	0.092	0.232	0.089	36
Non-game, News	0.263	0.090	0.241	0.076	0.282	0.086	0.247	0.069	36
Non-game, OHSUMED	0.196	0.098	0.207	0.080	0.228	0.090	0.217	0.085	36
Interface × Participant									
Game, Crowd	0.247	0.107	0.230	0.081	0.279	0.095	0.251	0.085	36
Game, Student	0.249	0.099	0.239	0.087	0.273	0.087	0.248	0.075	36
Non-game, Crowd	0.240	0.114	0.214	0.086	0.252	0.101	0.226	0.085	36
Non-game, Student	0.220	0.083	0.235	0.079	0.258	0.083	0.238	0.071	36
Collection × Participant									
News, Student	0262	0.084	0.246	0.081	0.278	0.084	0.247	0.066	36
News, Crowd	0.294	0.089	0.244	0.075	0.311	0.079	0.267	0.069	36
OHSUMED, Student	0.207	0.094	0.227	0.078	0.253	0.084	0.240	0.080	36
OHSUMED, Crowd	0.193	0.105	0.201	0.087	0.220	0.095	0.210	0.091	36

Table 65 Continued

	Query Refinement with Feedback								
	Precision@10		Avg. Precision		Recall		F-score		
Condition	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**N**
Interface × Collection × Participant									
Game, News, Student	0.295	0.072	0.256	0.088	0.301	0.075	0.264	0.059	18
Game, News, Crowd	0.290	0.095	0.242	0.073	0.313	0.084	0.270	0.074	18
Game, OHSUMED, Student	0.203	0.103	0.221	0.085	0.245	0.090	0.232	0.087	18
Game, OHSUMED, Crowd	0.204	0.102	0.220	0.089	0.246	0.097	0.232	0.092	18
Non-game, News, Student	0.228	0.082	0.236	0.075	0.255	0.089	0.229	0.069	18
Non-game, News, Crowd	0.299	0.086	0.247	0.079	0.310	0.076	0.264	0.066	18
Non-game, OHSUMED, Student	0.212	0.086	0.233	0.071	0.262	0.079	0.247	0.075	18
Non-game, OHSUMED, Crowd	0.180	0.110	0.181	0.082	0.194	0.089	0.187	0.085	18

7.10.2.1 Query Refinement with Feedback for P@10

Table 66: ANOVA results for query refinement with feedback for p@10

	Precision @ 10				
Effect	**SS**	**df**	**Mean Square**	**F**	**p-value**
Interface	0.012	1	0.012	1.398	0.239
Participant	0.003	1	0.003	0.349	0.556
Collection	0.022	1	0.022	25.498	0.000
Interface × Collection	0.004	1	0.004	0.492	0.484
Interface × Participant	0.004	1	0.004	0.500	0.481
Collection × Participant	0.020	1	0.020	2.332	0.129
Interface × Collection × Participant	0.026	1	0.026	3.067	0.082
Error	1.173	136	0.009		

Table 66 provides the ANOVA results for precision for the top 10 retrieved documents (p@10) for the Query Refinement with Feedback task. The three-way ANOVA

results for p@10 for the Query Refinement with Feedback task were statistically significant ($p < 0.05$) for the main effect of collection, with News having higher p@10 than the OHSUMED collection, $F(1,136)=25.498$, $p<0.001$. There were no statistically significant two-way or three-way interaction effects found for p@10 in Query Refinement with Feedback.

7.10.2.2 Query Refinement with Feedback AveP

Table 67 provides the ANOVA results for average precision (AveP) for the Query Refinement with Feedback task The three-way ANOVA results for AveP of Query Refinement with Feedback were statistically significant ($p < 0.05$) for the main effect of collection, with the News having higher AveP than the OHSUMED collection, $F(1,136)=12.511$, $p=0.001$. There were no statistically significant two-way or three-way interaction effects found for AveP in Query Refinement with Feedback.

Table 67: ANOVA results for query refinement with feedback average precision

Effect	Average Precision				
	SS	df	Mean Square	F	p-value
Interface	0.004	1	0.004	0.590	0.444
Participant	0.007	1	0.007	1.105	0.295
Collection	0.035	1	0.035	5.336	0.022
Interface × Collection	0.000	1	0.000	0.049	0.824
Interface × Participant	0.001	1	0.001	0.500	0.643
Collection × Participant	0.006	1	0.006	0.875	0.351
Interface × Collection × Participant	0.013	1	0.013	2.014	0.158
Error	0.884	136	0.007		

7.10.2.3 Query Refinement with Feedback Recall

Table 68: ANOVA results for query refinement with feedback recall

Effect	Recall				
	SS	df	Mean Square	F	p-value
Interface	0.015	1	0.015	2.131	0.147
Participant	0.000	1	0.000	0.000	0.996
Collection	0.121	1	0.121	16.691	0.000
Interface × Collection	0.000	1	0.000	0.067	0.796
Interface × Participant	0.001	1	0.001	0.204	0.652
Collection × Participant	0.040	1	0.040	5.572	0.020
Interface × Collection × Participant	0.029	1	0.029	4.009	0.047
Error	0.985	136	0.007		

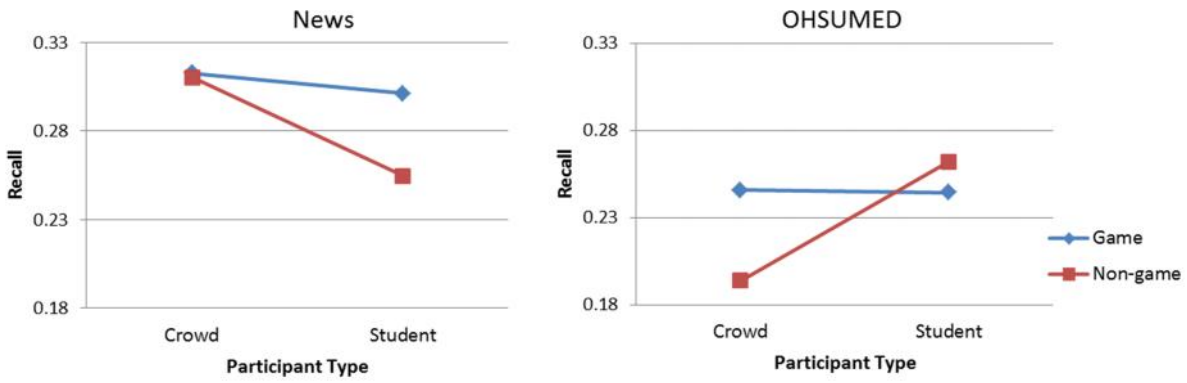

Figure 20: Interaction effects for recall in query refinement with feedback for interface (lines) and participant type (x-axis) for News collection (left) and OHSUMED collection (right)

Table 68 provides the ANOVA results for recall for the Query Refinement with Feedback task. The three-way ANOVA results for recall of Query Refinement with Feedback were statistically significant ($p < 0.05$) for the main effect of collection, with News having a higher recall than OHSUMED, $F(1,136)=16.691$, $p<0.001$ and the two-way interaction

collection x participant, F(1,136)=5.572, p=0.020. There was also a significant three-way interaction interface x participant type x collection interaction, F(1,136)=4.009, p=0.047. This three-way interaction effect is graphically depicted in Figure 20.

Simple effect follow-up tests for the three-way interaction using interface type revealed no significant effects. We then evaluated the interaction effects using collection type. This three-way interaction effect is graphically depicted in Figure 21.

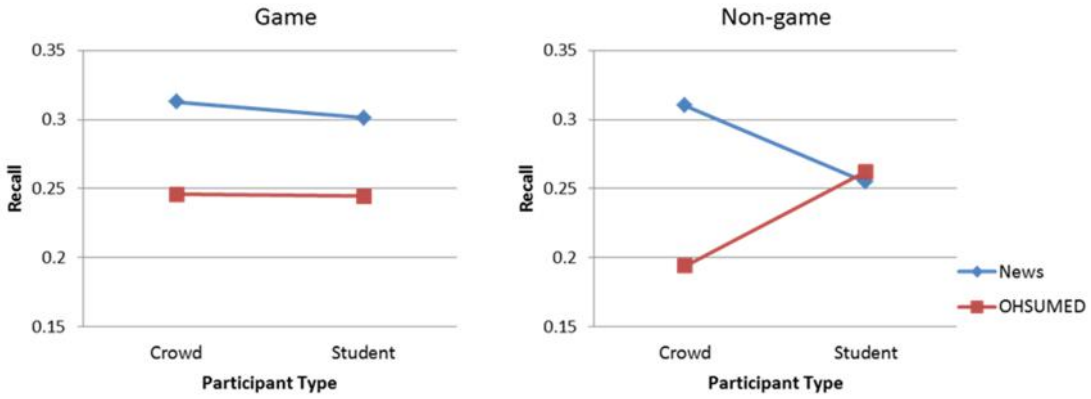

Figure 21: Interaction effects for recall in query refinement with feedback for interface (lines) and participant type (x-axis) for the game interface (left) and non-game interface (right)

Simple effect follow-up tests for the three-way interaction found that using the non-game interface with the News collection, crowd participants obtained higher recall than the students (M_C=0.310, M_S= 0.255; M_{Diff} = 0.056; t(136)=1.992; p=0.048; d=0.664), whereas with the OHSUMED collection students obtained higher recall than the crowd (M_S=0.262, M_C= 0.194; M_{Diff} = 0.68; t(136)=2.446; p=0.016; d=0.815).

7.10.2.4 Query Refinement with Feedback F-score

Table 69 provides the ANOVA results for F-score for the Query Refinement with Feedback task. The three-way ANOVA results for Query Refinement with Feedback F-score were statistically significant ($p < 0.05$) for the main effect of collection, with News having a

higher F-score than OHSUMED, F(1,136)=6.360, p=0.013. There were no statistically significant two-way or three-way interaction effects found for F-score in Query Refinement with Feedback.

Table 69: ANOVA results for query refinement with feedback for F-score

Effect	F-score				
	SS	**df**	**Mean Square**	**F**	**p-value**
Interface	0.011	1	0.011	1.930	0.167
Participant	0.001	1	0.001	0.135	0.714
Collection	0.037	1	0.037	6.360	0.013
Interface × Collection	0.000	1	0.000	0.044	0.834
Interface × Participant	0.002	1	0.002	0.352	0.554
Collection × Participant	0.023	1	0.023	3.859	0.052
Interface × Collection × Participant	0.018	1	0.018	3.019	0.085
Error	0.801	136	0.006		

7.10.3 Summary of Findings and Evaluation of Hypotheses

Table 70 provides a summary of our findings for the Initial Query task. Table 71 provides a summary of our findings for the Query Reformulation with Feedback task. From Tables 70 and 71, we can assess our hypotheses. This assessment is provided in Table 72.

Table 70: Summary of findings for the initial query task

Effect Type	Initial Query			
	Precision@10	**Ave Precision**	**Recall**	**F-score**
Main	Collection: News > OHSUMED; Interface: Game > Non-game	None	Collection: News > OHSUMED;	Collection: News > OHSUMED;
Interaction	Student & News: Game > Non-game	None	Non-game & News: Crowd > Students Non-game & OHSUMED: Students > Crowd	None

Table 71: Summary of findings for query reformulation with feedback task

Effect Type	Query Reformulation with Feedback			
	Precision@10	**Ave Precision**	**Recall**	**F-score**
Main	Collection: News > OHSUMED	Collection: News > OHSUMED	Collection: News > OHSUMED;	Collection: News > OHSUMED
Interaction	None	None	Non-game & News: Crowd > Students; Non-game & OHSUMED: Students > Crowd	None

Table 72: Analysis of hypotheses for the initial query and query reformulation with feedback tasks

Metric	Main Effect	Interaction Effects
Initial Query		
Precision@10	Reject H_O	Reject H_O
Average Precision	Do not reject H_O	Do not reject H_O
Recall	Reject H_O	Reject H_O
F-score	Reject H_O	Do not reject H_O
Query Refinement with Feedback		
Precision@10	Reject H_O	Do not reject H_O
Average Precision	Reject H_O	Do not reject H_O
Recall	Reject H_O	Reject H_O
F-score	Reject H_O	Do not reject H_O

7.11 Analysis

From the summary of findings (shown in Tables 70 and 71) and the analysis of our hypotheses (shown in Table 72), we can make some observations. The News collection provides higher scores than the OHSUMED collection across each of our four performance measures in both the Initial Query task and the Query Refinement with Feedback task. This better performance may be the result of a greater familiarity with the information need request in the News collection, whereas the more medically-inclined questions in OHSUMED were harder for participants to obtain phrases or use synonyms for terms in their queries.

Table 73 summarizes the relative performance for each topic in each collection. In the News collection, the game interface provides a higher p@10, average precision, recall and F-score for a majority of the 20 documents evaluated in both tasks. Comparing participant types, the crowd provides a higher p@10, recall, and F-score, but students are better at average precision for both tasks. In the OHSUMED collection, the game interface is better than the non-game interface across all four metrics in both tasks. Comparing

participant types, the students outperform the crowd across all four performance measures for both tasks on nearly all topics.

Table 73: Comparison of performance measurements for interface and participant types, by collection and task

Comparison	Initial Query				Query Reformulation w/Feedback			
	p@10	AveP	Recall	F-score	p@10	AveP	Recall	F-score
News								
Game > Non-game	20/20	14/20	19/20	15/20	20/20	14/20	19/20	17/20
Crowd > Students	18/20	4/20	20/20	13/20	19/20	9/20	20/20	14/20
OHSUMED								
Game > Non-game	15/20	18/20	18/20	18/20	13/20	18/20	19/20	19/20
Crowd > Students	0/20	0/20	0/20	0/20	2/20	0/20	0/20	0/20

7.11.1 Algorithmic approach vs. Human Computation Approach

As with our experiment in acronym identification and resolution, wish to examine the differences between the algorithmic approach and the human computation approaches we have empirically examined. We examine Average Precision (P), Recall (R) and F-score (F).

From Table 74, we can observe that the algorithm provides marginally better results than our human computation methods; the scores obtained using the algorithm are slightly above the median scores obtained during the original TREC-8 track using the same dataset (Hull & Robertson, 1999; S. E. Robertson & Walker, 1999) . This provides us with an opportunity to evaluate how human computation methods may augment the search process to improve results.

Table 74: Means and standard deviations of performance metrics for the mean human computation and algorithmic approaches.

Factor			Initial Query			Query Reformulation		
Participant	**Collection**	**Interface**	**P**	**R**	**F**	**P**	**R**	**F**
Crowd	News	Game	0.225	0.286	0.252	0.242	0.313	0.273
		Non-game	0.231	0.289	0.257	0.247	0.310	0.275
	OHSUMED	Game	0.211	0.232	0.221	0.220	0.246	0.232
		Non-game	0.174	0.190	0.182	0.181	0.194	0.188
Student	News	Game	0.244	0.280	0.261	0.256	0.301	0.277
		Non-game	0.223	0.231	0.227	0.236	0.255	0.245
	OHSUMED	Game	0.212	0.230	0.221	0.221	0.245	0.232
		Non-game	0.226	0.246	0.236	0.233	0.262	0.247
Algorithm	News	N/A	0.257	0.284	0.270	0.276	0.312	0.293
	OHSUMED	N/A	0.230	0.260	0.244	0.254	0.283	0.268

Observing from Table 74, if the algorithm could return a set of query results and we were able to somehow augment the results with the best the best performance from human computation methods, we could improve upon these methods as we did with the Acronym Resolution study. To accomplish this, we examine the first 3 performers in each collection from the crowd, based on F-score, and merge each of their query phrases with those supplied by the algorithm and remove duplicates. We examine if the improvement for each of our three performance metrics is statistically significant ($p<0.05$) in Table 75.

Table 75: Two-tailed t-test t-values and p-values for the comparison of the algorithm only and combined algorithm and human computation approaches for each metric for each collection

Task	Precision		Recall		F-score	
	T	**p**	**t**	**P**	**t**	**p**
Initial Query						
News	2.107	0.042*	3.972	< 0.001*	4.644	< 0.001*
OHSUMED	0.790	0.434	4.635	< 0.001*	2.976	0.005*
Query Reformulation with Feedback						
News	2.418	0.021*	2.046	0.048*	2.108	0.042*
OHSUMED	1.490	0.145	2.232	0.032*	2.328	0.025*

From Table 75, we see that in the News collection across our three metrics, the combined approach provides a significant improvement over the algorithmic approach alone. In the OHSUMED collection we obtain a significant improvement for recall and F-score. As compared with earlier algorithm only submissions in earlier TREC campaigns, these findings illustrate the merits of an augmented human computation approach to query formulation.

Examining Table 75, for each collection, the combined set of terms from the algorithmic approach and first 3 human computation participants provides an improvement over either approach alone. In addition, the number of unique terms increases in the combined approach – in some situations, it nearly doubles. Although this increase in the number of terms improves recall as expected, in the News collection, the precision score also improved significantly. This indicates the set of terms provided by the combined approach improves recall but not at the expense of precision – one of the positive outcomes from this augmented approach.

7.11.2 Evaluation of search phrases

In 14.7% of the queries refined after receiving feedback that we observed, the F-score actually dropped. In most cases, this was due to a reduction in precision as new terms were added that do not aid the query. We would like to determine which participant and interface types provide for the largest increase in the F-score between the Initial Query (IQ) and the Query Refinement with Feedback (QR), as may indicate the method can provide response to the additional feedback provided. Table 76 provides this information for the mean F-scores for each of our factors in each collection.

Table 76: Comparison of F-score for mode of interaction and participant types, by collection and task

Collection	Factor	Initial Query	Query Refinement	% Improvement
OHSUMED	Game	0.221	0.232	5.2%
	Non-game	0.200	0.217	8.5%
News	Game	0.257	0.275	7.0%
	Non-game	0.242	0.260	7.5%
OHSUMED	Crowd	0.201	0.210	4.3%
	Student	0.228	0.240	5.0%
News	Crowd	0.254	0.274	7.7%
	Student	0.244	0.261	6.9%

From Table 76, we observe that for both collections, the non-game provides a higher increase in the F-score. This may be a result of less pressure in the non-game interface allowing for better concentration; the game interface has a time constraint that pressures participants to enter terms quickly and has more distractions (e.g., music, countdown timer) that may affect the concentration of participants when enhancing to their original query terms. Even with this pressure, the difference in F-score increases between the two interface types was slight.

The increase in F-score due to participant type gave a more mixed picture. The overall improvement for participants was greater between phases in the News collection.

Crowd participants improved their F-scores by a greater percentage than students in the News collection. In the OHSUMED collection, however, it was students improved their F-scores by a greater measure than the crowd. The greater improvement in each collection between phases may be due to the greater diversity, or number of unique non-stopword phrases provided by the crowd for News and students for OHSUMED, which is provided in Table 77.

Table 77: Number of unique non-stopword phrases across all 20 topics provided by each participant type by collection.

Collection	Participant Type	Initial Query	Query Refinement	% Change
OHSUMED	Crowd	95	104	9.5%
	Student	102	122	19.6%
News	Crowd	138	166	20.3%
	Student	131	146	11.5%

In Table 77, we see that the participant types that provide a larger number of terms in the Query Refinement with Feedback phase also obtained a larger increase in F-score, which reinforces the information found in Table 76. Therefore, the quality of feedback is likely important to the query refinement – providing more information, such as positive and negative examples of successful documents is likely to improve the results even further.

We observed that the combination of terms generated by the algorithm along with the terms provided by human computation participants increase the final precision and F-scores to a value considerably higher than the algorithmic approach alone. As with our other experiments, this indicates that we can obtain quality improvements with as few as three crowdworkers. Again, this finding reinforces our earlier hypothesis that augmenting algorithmic approaches can provide substantial value to tasks such as query refinement.

7.12 Conclusion

One of the fundamental steps in IR is formulating queries to find relevant documents. Along with this is the refinement of queries once an initial query result has been returned by the search engine. In this study, we examine the performance of several factors of human computation including mode of interaction (using a game interface and a non-game interface) and type of participant (students and crowdworkers). We evaluate each combination in two collections (News collection and the OHSUMED collection, which is comprised of medical documents) on four metrics, p@10, average precision, recall, and F-score. Additionally, we examine how these human computation factors compare with an algorithm that has performed very well in past TREC query retrieval experiments. Our human participants did not outperform the algorithm in either data collection, even when using the same process to retrieve results.

We find that the News collection provides higher performance across our measures as compared with the OHSUMED collection, as expected. Overall, game interfaces provide higher performance metrics than non-game interfaces, and there is no noticeable difference between crowd participants and students. We also find that when using the non-game interface, the crowd outperforms students in the News collection; students outperform the crowd for the OHSUMED collection. The difference in interfaces may relate to the difficulty factor of the task – tedious repetitive tasks appear to be better designed for using a game interface.

Improved performance was tied to the number of non-stoplisted terms provided in both the Initial Query and Query Refinement with Feedback tasks. Therefore we believe that collaborative human computation methods that can combine user-supplied phrases and synonyms may increase performance considerably.

CHAPTER 8

APPLICATION OF OUR HUMAN COMPUTATION FRAMEWORK TO OTHER AREAS OF INFORMATION SCIENCE

Although much of this thesis has focused on applying the human computation framework developed in Chapter 3 to Information Retrieval. In his 1968 article, "Information Science: What Is It?" Howard Borko summarizes information science as follows (Borko, 1968).

> It is an interdisciplinary science that investigates the properties and behavior of information, the forces that govern the flow and use of information, and the techniques, both manual and mechanical, of processing information for optimal storage, retrieval, and dissemination.

A more updated definition of information science is "a an interdisciplinary field primarily concerned with the analysis, collection, classification, manipulation, storage, retrieval, movement, and dissemination of information" (Stock & Stock, 2013). Many of the information science tasks mentioned in the updated definition parallel Steps in our IR model with broader implications.

Analysis of information, broadly defined, is the process of breaking complex information into smaller components in order to obtain a better understanding. It involves inspecting, cleaning, transforming, and modeling the underlying data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making, the type of tasks that humans can do well. Since these analysis tasks involve the transformation of data into knowledge, it appears most suitable for human involvement. Not only does this type of information analysis have multiple facets and approaches, but it requires the employment of a diverse set of techniques, many of which the framework provided in Chapter 3 has addressed. More importantly, human computation methods, if designed properly, can perform this analysis quickly, inexpensively, and accurately.

Examining our framework, our first criteria examine the scale of the task and the need for expert assistance with the task. The rapid development of "big data" analysis tools in the past decade has made the scalability of large amounts of data more achievable. Expert guidance in approaches and techniques from software has also become readily available. For many information analysis tasks, crowd or game assistance is a viable option, particularly in highlighting useful information or decision support, but also inspecting, cleaning, transforming and modeling data which can be augmented by the human computation approaches discussed in this thesis.

Some challenges remain, such as the significant amount of information that requires keen attention to privacy and anonymity policies; few companies who are contractually required to protect information will allow anonymous participants from the crowd to provide analysis on this data. This may be the biggest obstacle to permitting crowd participation in many data analysis tasks, and requires companies with this privacy need to keep this activity in a controllable environment.

Criterion 3 from our framework could be met provided the information could be integrated into an overall process. Additionally, as Criteria 4 and 5 indicate, it is possible to design an information analysis task that generates useful analysis while enriching the experience and encouraging game play. Several well-known games that involve the analysis of information, such as TagATune and the ESP Game, demonstrate that information analysis is indeed possible to achieve through the game or a crowdsourcing platform.

Collecting information is similar to the second Step of our IR model, which involves setting up a collection, but again, with broader implications. The task of information collection would benefit from the diversity of the crowd. This diversity may provide guidance on what information is available, how it could be presented, and where it can be obtained; the categorization of information, in which ideas and objects are recognized, differentiated, and understood, also appears to be a suitable task according to our framework.

Categorization implies that objects are grouped into categories, usually for some specific purpose. Ideally, a category illuminates a relationship between the subjects and objects of knowledge. Therefore, categorization is a task that likely can scale, can be integrated and it is possible to make this task interesting through a game format, satisfying the first 5 criteria of our framework. Both collecting and categorizing information tasks can be augmented by the crowd improve the results, much as we found was possible in Chapter 5 with our experiment on acronym resolution.

The manipulation and storage of information follows many of the pre-processing Steps and the indexing Step (Steps 3 and 4 of our IR model). This task is frequently associated with databases or search engines and therefore the discussion about how our framework applies to the preprocessing and indexing Steps would apply to these tasks as well.

Overall, the role of human computation methods in information science is promising, given our experiments here. This is particularly true for those tasks in which humans can augment Algorithmic processes, allowing humans to apply their understanding of what humans expect and provide the difficult decisions, while allowing the algorithm to handle the scale of large tasks.

CHAPTER 9

CONCLUSION

The studies described in this thesis help to investigate the following three research questions we raised in the Introduction chapter of this thesis. First, we wish to see if we can use human computation to improve performance in Information Science (IS) tasks, particularly those involving Information Retrieval (IR). Second, we wish to determine which human computation factors are most appropriate for a given IS task. Finally, we wish to determine in which situations the augmentation of an algorithmic approach with human computation makes sense.

We began our investigation by developing a process-based IR model. Not only does the IR model represent an important, algorithm-intensive area of IS, but it also can be broken into a number of discrete steps. This provides us with an opportunity to apply our framework to each step independently. Using this, we evaluate our framework criteria and identified three steps that were candidates for further study.

For our first experiment, we examined acronym identification and resolution. Our framework initially indicated would not scale well to use human computation, but another aspect of our framework indicated it might be an excellent candidate for using human computation to augment augmentation. There have been a number of algorithm-based acronym identification and resolution tools that claim strong recall and precision, but nearly all are specific to one domain. We avoid algorithms that are lexicon dependent and those that require advance training because of their reliance on the single domain. We examine a rule-based algorithm that has done well in identifying and resolving acronyms in medical text and apply it to two new domains – a general collection of News documents and a specialized collection of Patents. To our knowledge, there have been no other studies that evaluate the role of human computation in acronym resolution and detection.

In addition to two different stages to finding definitions of algorithms (identification and resolution), and two different collections (News and Patents), we examine the mode of interaction and the type of participant. Given the recent rise in attention on GWAPs and gamification, examining the mode of interaction allows us to see if Games With A Purpose (GWAP) are more effective than non-game web interfaces for different IR steps. We also compare mostly anonymous workers from a crowdsourcing platform with students, which are traditionally used in studies. This allows us to examine if crowdworkers can become an inexpensive proxy to a traditional method of conducting IR research. We apply these to the News collection we examined in the acronym study but also to the OHSUMED collection, which is comprised of documents containing medical abstracts. Overall, we find that human computation outperforms the algorithm for precision and F-score, but not for recall. We also find that game interfaces work best when tasks are mundane and require little deep concentration, as we discovered with in the identification task using the less-challenging News collection. Students outperformed the crowd in several scenarios, particularly those involving more external common knowledge, such as resolving definitions in the Patent collection. More importantly, we find that augmented processes, where the algorithm finds as many acronyms as it can then turns it over to human computation to use and apply external common knowledge, can substantially improve the results for identification and resolution.

Our second experiment focused on evaluating the relevance of documents in a collection to an information need. Human assessment costs are the most expensive aspect of relevance judgments, and a pooling process is usually employed to reduce this cost. We wished to examine if an augmented process using algorithms and human computation could further reduce the cost of pooling while still maintaining quality across four performance measures: precision, recall, F-score, and LAM, a misclassification rate normally used in spam detection. We examined two algorithms, one which applies a ranking algorithm derived from the submissions of others, and another that uses text clustering to group documents with

similar terms together. An examination of these algorithms across these two different collections finds that the non-clustering method improves precision, F-score and LAM, while clustering improves recall.

We then conduct a deeper investigation on the News collection, which reinforces our findings on our performance measures for each algorithm. We find assessment costs are substantially lower than the costs associated with the pooling process. Even if we conservatively assume the same assessment costs per document for pooling, we find that our methods are 100% to 1600% cheaper; when using the estimated true costs, our methods are hundreds of times less expensive. Thus, our augmented human computation-algorithm approaches are able to significantly reduce the assessment costs while only marginally reducing the quality across our performance measures.

Our third experiment examined another IR step that our framework indicates would be a good application of human computation: obtaining a query for an information need and making refinements to that query upon receiving some feedback. The same modes of interaction and the same participant types were examined as in our acronym experiment, but with the two collections used in the relevance assessment experiment. We compared these human computation methods against an algorithm that has demonstrated strong performance in several TREC query retrieval campaigns. Although the algorithm outperformed our human computation methods, when we augmented the terms provided by algorithm with those provided by the first three crowdworkers to participate, we obtain statistically significant increase over the algorithm alone across all performance metrics in the News collection and for recall and F-score in the OHSUMED collection.

Therefore, in all three of the IS tasks we examined, we observed human computation can improve upon existing processes, either directly, as was found in precision during acronym resolution, or indirectly through an augmented method, as was found in the relevance assessment algorithms that improved upon existing pooling processes. The human computation approach that is most appropriate depends on several factors, including the level

of difficulty of the task (lower difficulty favors the crowd and game interfaces, higher difficulty slightly favors students and non-game interfaces), how tasks can effectively scale, how entertaining the task can be made, and how well traditional algorithms can perform the same task. In each of the studies examined in this thesis, augmented approaches, which have not been examined much in the literature, can provide the most effective solution. However, they need to be designed so the human and machine components can integrate well. Additionally, there must be a mechanism to indicate which tasks the machine cannot perform and require human intervention. With the acronym resolution, for example, our algorithm was able to resolve most acronyms, which allows it to scale, but it was also able to mark those acronyms it could not identify with confidence. Humans could then address the resolution of the few, more challenging acronyms.

In summary, we have examined three different modes of interaction (algorithm, non-game interfaces, and game interfaces), two different participant types (crowdworkers and students/casual users), and two different types of document collections. Each of these factors has presented their own merits: for participants, algorithms provide scale, the crowd can provide diversity and quality at a low price, and students provide a well-studied educated demographic that can adequately perform a great number of tasks. Likewise, in our study, game interfaces have demonstrated their ability to improve performance, particularly on tedious tasks that do not require deep concentration. Non-game interfaces have shown they are best when time and user concentration are paramount. However, the best performance is possible when we can apply the strengths of each factor, such as using humans to augment algorithm processes in acronym resolution or query formulation. Our framework and our experiments have shown that future quality improvements in Information Science are indeed possible using human computation, and in particular using an augmented approach.

REFERENCES

Abt, C. C. (1987). Serious games: Univ Pr of Amer.

Ageev, M., Guo, Q., Lagun, D., & Agichtein, E. (2011). Find it if you can: a game for modeling different types of web search success using interaction data. SIGIR (Vol. 11, pp. 345-354).

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 37-45). ACM.

Almquist, B., Mejova, Y., Ha-Thuc, V., & Srinivasan, P. (2008). University of Iowa at TREC 2008 Legal and Relevance Feedback Tracks: DTIC Document.

Alonso, O. (2011). Perspectives on Infrastructure for Crowdsourcing. Crowdsourcing for Search and Data Mining (CSDM 2011), 7.

Alonso, O., & Baeza-Yates, R. (2011). Design and implementation of relevance assessments using crowdsourcing. Advances in information retrieval (pp. 153-164). Springer Berlin Heidelberg.

Alonso, O., & Lease, M. (2011a). Crowdsourcing 101: putting the WSDM of crowds to work for you. WSDM (pp. 1-2).

Alonso, O., & Lease, M. (2011b). Crowdsourcing for information retrieval: principles, methods, and applications. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (pp. 1299-1300). ACM.

Alonso, O., & Mizzaro, S. (2009). Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation (pp. 15-16).

Alonso, O., & Mizzaro, S. (2012). Using crowdsourcing for TREC relevance assessment. Information Processing & Management, 48(6), 1053-1066. doi: 10.1016/j.ipm.2012.01.004

Anick, P. (2003). Using terminological feedback for web search refinement: a log-based study. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 88-95). ACM.

Araujo, R. M. (2013). 99designs: An Analysis of Creative Competition in Crowdsourced Design. Conference on Human Computation & Crowdsourcing (HCOMP'13)

Ashton, A. H. (1985). Does consensus imply accuracy in accounting studies of decision making? Accounting Review, 173-185.

Aurenhammer, F. (1991). Voronoi diagrams—a survey of a fundamental geometric data structure. ACM Computing Surveys (CSUR), 23(3), 345-405.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 82): Addison-Wesley New York.

Baumeister, J., Reutelshoefer, J., & Puppe, F. (2007). Knowwe: community-based knowledge capture with knowledge wikis. Proceedings of the 4th international conference on Knowledge capture (pp. 189-190). ACM.

Belkin, N. J., Cool, C., Kelly, D., Lin, S.-J., Park, S. Y., Perez-Carballo, J., & Sikora, C. (2001). Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. Inf. Process. Manage., 37(3), 403-434. doi: 10.1016/s0306-4573(00)00055-8

Bell, M., Reeves, S., Brown, B., Sherwood, S., MacMillan, D., Ferguson, J., & Chalmers, M. (2009). Eyespy: Supporting navigation through play. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 123-132). ACM.

Bennett, P. N., Chickering, D. M., & Mityagin, A. (2009a). Learning consensus opinion: mining data from a labeling game. Proceedings of the 18th international conference on World wide web (pp. 121-130). ACM.

Bennett, P. N., Chickering, D. M., & Mityagin, A. (2009b). Picture this: preferences for image search. Proceedings of the 18th international conference on World wide web (pp. 121-130). ACM.

Berto, A., Mizzaro, S., & Robertson, S. (2013). On Using Fewer Topics in Information Retrieval Evaluations. Proceedings of the 2013 Conference on the Theory of Information Retrieval (p. 9). ACM.

Bigham, J. P., Kaminsky, R. S., Ladner, R. E., Danielsson, O. M., & Hempton, G. L. (2006). WebInSight:: making web images accessible. Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility (pp. 181-188). ACM.

Bingham, A., & Spradlin, D. (2011). The Open Innovation Marketplace: Creating Value in the Challenge Driven Enterprise: Ft Pr.

Blanco, R., Halpin, H., Herzig, D. M., Mika, P., Pound, J., Thompson, H. S., & Duc, T. T. (2011). Repeatable and reliable search system evaluation using crowdsourcing. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (pp. 923-932). ACM.

Bonabeau, E. (2009). Decisions 2.0: The power of collective intelligence. MIT Sloan Management Review, 50(2), 45-52.

Borko, H. (1968). Information science: What is it? American Documentation, 19(1), 3-5. doi: 10.1002/asi.5090190103

Bozzon, A., Brambilla, M., & Ceri, S. (2012). Answering search queries with CrowdSearcher. Proceedings of the 21st international conference on World Wide Web (pp. 1009-1018). ACM.

Bradley, P. S., & Fayyad, U. M. (1998). Refining initial points for k-means clustering. ICML (Vol. 98, pp. 91-99).

Brenes, D. J., Gayo-Avello, D., P, K., Perez-Gonzalez. (2009). Survey and evaluation of query intent detection methods. Proceedings of the 2009 workshop on Web Search Click Data (pp. 1-7). ACM.

Buttcher, S., Clarke, C. L. A., & Cormack, G. V. (2004). Domain-specific synonym expansion and validation for biomedical information retrieval (MultiText experiments for TREC 2004). Paper presented at the Proceedings of the 13th Text Retrieval Conference.

Carvalho, V. R., Lease, M., & Yilmaz, E. (2011). Crowdsourcing for search evaluation. ACM Sigir forum (Vol. 44, No. 2, pp. 17-22). ACM.

Casey, S., Kirman, B., & Rowland, D. (2007). The gopher game: a social, mobile, locative game with user generated content and peer review. Proceedings of the international conference on Advances in computer entertainment technology (pp. 9-16). ACM.

Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcase, M., & Poesio, M. (Eds.). (2013). Using Games to Create Language Resources: Successes and Limitations of the Approach. Berlin: Springer.

Chan, K. T., King, I., & Yuen, M. C. (2009). Mathematical modeling of social games. Computational Science and Engineering, 2009. CSE'09. International Conference on (Vol. 4, pp. 1205-1210). IEEE.

Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. Journal of Economic Behavior & Organization.

Chklovski, T., & Gil, Y. (2005). Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors. Proceedings of the 3rd international conference on Knowledge capture (pp. 35-42). ACM.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 659-666). ACM.

Cleverdon, C. (1967). The Cranfield tests on index language devices. Paper presented at the Aslib proceedings. Aslib proceedings (Vol. 19, No. 6, pp. 173-194). MCB UP Ltd.

Croft, W. B., Metzler, D., & Strohman, T. (2010). Search engines: Information retrieval in practice: Addison-Wesley.

Csikszentmihalyi, M. (1991). Flow: The psychology of optimal experience: Harper Perennial.

Dannélls, D. (2006). Automatic acronym recognition. Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations (pp. 167-170). Association for Computational Linguistics.

Darpa Network Challenge. (2009), from http://archive.darpa.mil/networkchallenge/

Dasdan, A., Drome, C., Kolay, S., Alpern, M., Han, A., Chi, T., . . . Verma, S. (2009). Thumbs-Up: a game for playing to rank search results. Paper presented at the Proceedings of the ACM SIGKDD Workshop on Human Computation, Paris, France.

Davies, A. (2009). Crowdsourcing News: The Guardian and MP expenses Retrieved April 3, 2012, from http://idioplatform.com/2009/06/crowdsourcing-news-the-guardian-and-mp-expenses/

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1-38.

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011). Gamification. using game-design elements in non-gaming contexts. Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems (pp. 2425-2428). ACM.

Dillon, A., & Song, M. (2006). An empirical comparison of the usability for novice and expert searchers of a textual and a graphic interface to an art-resource database. Journal of Digital Information, 1(1).

Efthimiadis, E. N. (2000). Interactive query expansion: a user-based evaluation in a relevance feedback environment. J. Am. Soc. Inf. Sci., 51(11), 989-1003. doi: 10.1002/1097-4571(2000)9999:9999<::aid-asi1002>3.0.co;2-b

Eickhoff, C., de Vries, A. (2011). How crowdsourcable is your task. In Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the fourth ACM international conference on web search and data mining (WSDM) (pp. 11-14).

Eickhoff, C., Harris, C. G., de Vries, A. P., & Srinivasan, P. (2012). Quality through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments. Paper presented at the SIGIR'12, Portland, OR.

Eickhoff, C., Harris, C. G., Srinivasan, P., & de Vries, A. P. (2011). GeAnn at the TREC 2011 Crowdsourcing Track.

Ekins, S., & Williams, A. (2010). Reaching Out to Collaborators: Crowdsourcing for Pharmaceutical Research. Pharmaceutical Research, 27(3), 393-395. doi: 10.1007/s11095-010-0059-0

Farkas, D. F. (2003). Food engineering history. Encyclopedia of Agricultural Food and Biological Engineering, New York: Marcel Dekker, 346-349.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. Knowledge Discovery and Data Mining, 82-88.

Feng, S., Xiong, Y., Yao, C., Zheng, L., & Liu, W. (2009). Acronym extraction and disambiguation in large-scale organizational web pages. Proceedings of the 18th ACM conference on Information and knowledge management (pp. 1693-1696). ACM.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., & Dredze, M. (2010). Annotating named entities in Twitter data with crowdsourcing. Paper presented at the Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, California.

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon mechanical turk: Gold mine or coal mine? Computational Linguistics, 37(2), 413-420.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. The computer journal, 41(8), 578-588.

Franklin, M., Kossmann, D., Kraska, T., Ramesh, S., & Xin, R. (2011). CrowdDB: Answering queries with crowdsourcing. Proc. SIGMOD (2011), 61-72.

Gartner. (2011). Gartner says by 2015, more than 50 percent of organization that manage innovation processes will gamify those processes. http://www.gartner.com/newsroom/id/1629214

Grady, C., & Lease, M. (2010). Crowdsourcing document relevance assessment with Mechanical Turk.

Grant, L., Daanen, H., Benford, S., Hampshire, A., Drozd, A., & Greenhalgh, C. (2007). MobiMissions: the game of missions for mobile phones. ACM SIGGRAPH 2007 educators program (p. 12). ACM.

Grier, D. A. (2005). When computers were human: Princeton Univ Pr.

Groh, F. (2012). Gamification: State of the Art Definition and Utilization. Institute of Media Informatics Ulm University, 39.

Hacker, S., & Von Ahn, L. (2009). Matchin: eliciting user preferences with an online game. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1207-1216). ACM

Hagen, J. W., & Kingsley, P. R. (1968). Labeling effects in short-term memory. Child development, 113-121.

Hamari, J. (2013). Transforming Homo Economicus into Homo Ludens: a field experiment on gamification in a utilitarian peer-to-peer trading service. Electronic Commerce Research and Applications.

Hamari, J., Koivisto, J., & Sarsa, H. (2014) Does Gamification Work?—A Literature Review of Empirical Studies on Gamification. Procedings of the 47th Hawaii International Confernece on System Sciences, Hawaii USA (to appear)

Harris, C., & Srinivasan, P. (2012). Using Hybrid Methods for Relevance Assessment in TREC Crowd'12. Online Proceedings of TREC.

Harris, C. G. (2012). An Evaluation of Search Strategies for User-Generated Video Content. CrowdSearch (pp. 48-53)

Harris, C. G., & Srinivasan, P. (2012). Applying Human Computation Mechanisms to Information Retrieval. Proceedings of the American Society for Information Science and Technology, 49(1), 1-10.

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 203-212). ACM.

Hellerstein, J. M., & Tennenhouse, D. L. (2011). Searching for Jim Gray: a technical overview. Communications of the ACM, 54(7), 77-87.

Hersh, W., & Bhupatiraju, R. T. (2003). TREC genomics track overview. TREC 2003.

Hersh, W., Buckley, C., Leone, T., & Hickam, D. (1994). OHSUMED: an interactive retrieval evaluation and new large test collection for research. SIGIR'94 (pp. 192-201). Springer London.

Ho, C. J., Chang, T. H., Lee, J. C., Hsu, J. Y., & Chen, K. T. (2009). KissKissBan: a competitive human computation game for image annotation. Proceedings of the ACM SIGKDD Workshop on Human Computation (pp. 11-14). ACM.

Hosseini, M., Cox, I. J., Milić-Frayling, N., Kazai, G., & Vinay, V. (2012). On aggregating labels from multiple crowd workers to infer relevance of documents Advances in Information Retrieval (pp. 182-194): Springer.

Howe, J. (2006). The Rise of Crowdsourcing. Wired magazine, 14(6), 1-4.

Howe, J. (2008). Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business. New York, NY: Crown Publishing Group.

Hu, C., Bederson, B. B., Resnik, P., & Kronrod, Y. (2011). Monotrans2: A new human computation system to support monolingual translation. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1133-1136). ACM.

Hull, D. A., & Robertson, S. E. (1999). The TREC-8 Filtering Track Final Report. In TREC.

Huotari, K., & Hamari, J. (2012). Defining gamification: a service marketing perspective. Proceeding of the 16th International Academic MindTrek Conference (pp. 17-22). ACM.

Ipeirotis, P., Chandrasekar, R., & Bennett, P. (2010). Report on the human computation workshop. HCOMP 2010.

Ipeirotis, P. G., Chandrasekar, R., & Bennett, P. (2010). A report on the human computation workshop (HComp 2009). SIGKDD Explor. Newsl., 11(2), 80-83. doi: 10.1145/1809400.1809416

Ipeirotis, P. G., & Paritosh, P. K. (2011). Managing crowdsourced human computation: a tutorial. Proceedings of the 20th international conference companion on World wide web (pp. 287-288). ACM.

Ipeirotis, P. G., Provost, F., & Wang, J. (2010). Quality management on amazon mechanical turk. Proceedings of the ACM SIGKDD workshop on human computation (pp. 64-67). ACM.

Joachims, T. (1996). A Probabilistic Analysis of the Rocchio Algorithm with TFI*DF* for Text Categorization: No. CMU-CS-96-118. CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE, 1996.

Kamar, E., Hacker, S., & Horvitz, E. (2012). Combining human and machine intelligence in large-scale crowdsourcing. In Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1 (pp. 467-474). International Foundation for Autonomous Agents and Multiagent Systems.

Kazai, G., Milic-Frayling, N., & Costello, J. (2009). Towards methods for the collective gathering and quality control of relevance assessments. Paper presented at the Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, Boston, MA, USA.

Kinsbourne, M., & George, J. (1974). The mechanism of the word-frequency effect on recognition memory. Journal of Verbal Learning and Verbal Behavior, 13(1), 63-69.

Kokkinakis, D., & Dannélls, D. (2006). Recognizing acronyms and their definitions in Swedish medical texts. Proceedings of the 5th Conference on Language Resources and Evalutaion (LREC). Genoa, Italy.

Kumar, A., & Lease, M. (2011). Learning to rank from a noisy crowd. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 1221-1222. ACM, 2011.

Lancaster, F. W., & Warner, A. J. (1993). Information Retrieval Today (1 ed.). Arlington, VA, USA: Information Resources Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. biometrics, 159-174.

Larkey, L. S., Ogilvie, P., Price, M. A., & Tamilio, B. (2000). Acrophile: an automated acronym extractor and server. In Proceedings of the fifth ACM conference on Digital libraries (pp. 205-214). ACM.

Law, E., Ahn, L. v., & Mitchell, T. (2009). Search war: a game for improving web search. Proceedings of the ACM SIGKDD Workshop on Human Computation (pp. 31-31). ACM.

Law, E., & Von Ahn, L. (2009). Input-agreement: a new mechanism for collecting data using human computation games. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1197-1206). ACM.

Lawrence, K. F. (2011). Crowdsourcing Linked Data From Shakespeare To Dr Who. Proceedings of Web Science (2011).

Lease, M., & Yilmaz, E. (2012). Crowdsourcing for information retrieval. ACM SIGIR Forum (Vol. 45, No. 2, pp. 66-75). ACM.

Lévy, P. (1999). Collective Intelligence: Mankind's Emerging World in Cyberspace (R. Bonomo, Trans.). New York, NY: Perseus Publishing

Lieberman, H., Smith, D., & Teeters, A. (2007). Common Consensus: a web-based game for collecting commonsense goals. ACM Workshop on Common Sense for Intelligent Interfaces. 2007.

Liu, T. X., Yang, J., Adamic, L. A., & Chen, Y. (2011). Crowdsourcing with All-pay Auctions: a Field Experiment on Taskcn. Proceedings of the American Society for Information Science and Technology 48.1 (2011): 1-4.

Luon, Y., Aperjis, C., & Huberman, B. A. Rankr: A Mobile System for Crowdsourcing Opinions. Mobile Computing, Applications, and Services. Springer Berlin Heidelberg, 2012. 20-31.

Łupkowski, P. (2011). Human computation–how people solve difficult AI problems (having fun doing it). Homo Ludens, 1(3).

Ma, H., Chandrasekar, R., Quirk, C., & Gupta, A. (2009). Improving search engines using human computation games. Proceedings of the 18th ACM conference on Information and knowledge management (pp. 275-284). ACM.

Malone, T. W. (1980). What makes things fun to learn? Heuristics for designing instructional computer games. Proceedings of the 3rd ACM SIGSMALL symposium and the first SIGPC symposium on Small systems (pp. 162-169). ACM.

Malone, T. W. (1982). Heuristics for designing enjoyable user interfaces: Lessons from computer games. Proceedings of the 1982 conference on Human factors in computing systems (pp. 63-68). ACM.

Malone, T. W., Laubacher, R., & Dellarocas, C. N. (2009). Harnessing crowds: Mapping the genome of collective intelligence: MIT Sloan.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). Introduction to information retrieval (Vol. 1): Cambridge University Press Cambridge.

Mataric, M. J. (1993). Designing emergent behaviors: From local interactions to collective intelligence. Proceedings of the Second International Conference on Simulation of Adaptive Behavior (pp. 432-441).

Matyas, S. (2007). Playful geospatial data acquisition by location-based gaming communities. The International Journal of Virtual Reality, 6(3), 1-10.

Matyas, S., Matyas, C., Schlieder, C., Kiefer, P., Mitarai, H., & Kamata, M. (2008). Designing location-based mobile games with a purpose: collecting geospatial data with CityExplorer. Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology (pp. 244-247). ACM.

McCann, R., Doan, A., Varadarajan, V., Kramnik, A., & Zhai, C. (2003). Building data integration systems: A mass collaboration approach. IIWeb (pp. 183-188).

McCreadie, R., Macdonald, C., & Ounis, I. (2013). Identifying top news using crowdsourcing. Information Retrieval, 1-31.

McCreadie, R. M. C., Macdonald, C., & Ounis, I. (2010). Crowdsourcing a news query classification dataset. Proceedings of the ACM SIGIR 2010 workshop on crowdsourcing for search evaluation (CSE 2010) (pp. 31-38).

McGonigal, J. (2011). Reality is broken: Why games make us better and how they can change the world: Penguin Pr.

McKibbon, K. A., Haynes, R. B., Walker Dilks, C. J., Ramsden, M. F., Ryan, N. C., Baker, L., . . . Fitzgerald, D. (1990). How good are clinical MEDLINE searches? A comparative study of clinical end-user and librarian searches. Computers and Biomedical Research, 23(6), 583-593. doi: 10.1016/0010-4809(90)90042-b

Metzler, D., Strohman, T., & Croft, W. B. (2006). Indri at trec 2006: Lessons learned from three terabyte tracks: DTIC Document.

Metzler, D., Strohman, T., Turtle, H., & Croft, W. B. (2004). Indri at TREC 2004: Terabyte track: DTIC Document.

Meyer, P., Hoeng, J., Norel, R., Sprengel, J., Stolle, K., Bonk, T., . . . Rice, J. J. (2012). Industrial Methodology for Process Verification in Research (IMPROVER): Towards Systems Biology Verification. Bioinformatics.

Michaelsen, L. K., Watson, W. E., & Black, R. H. (1989). A realistic test of individual versus group consensus decision making. Journal of Applied Psychology, 74(5), 834.

Milne, D., Nichols, D. M., & Witten, I. H. (2008). A competitive environment for exploratory query expansion. Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (pp. 197-200). ACM.

Mollov, M. (2010). AcronymFinder.com passes the 1 million definition milestone. Retrieved from http://blog.acronymfinder.com/2010/12/acronymfindercom-passes-1-million.html

Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Searching Microblogs: Coping with Sparsity and Document Quality. Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 183-188). ACM.

Nowak, S., & Rüger, S. (2010). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. Proceedings of the international conference on Multimedia information retrieval (pp. 557-566). ACM.

O'Neil, C., Purvis, J., & Azzopardi, L. (2011). Fu-Finder: a game for studying querying behaviours. Proceedings of the 20th ACM international conference on Information and knowledge management, Glasgow, Scotland, UK.

Odumuviwa, V., & David, A. (2009). A user centered approach to collaborative information retrieval. Collaborative Technologies and Systems, 2009. CTS'09. International Symposium on (pp. 494-501). IEEE.

Oostdijk, N., Verberne, S., & Koster, C. (2010). Constructing a broad-coverage lexicon for text mining in the patent domain. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010). European Language Resources Association (ELRA).

Paiement, J. F., Shanahan, J. G., & Zajac, R. (2010). Crowd Sourcing Local Search Relevance. Collaborative Technologies and Systems, 2009. CTS'09. International Symposium on (pp. 494-501). IEEE.

Parent, G., & Eskenazi, M. (2011). Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges. INTERSPEECH (pp. 3037-3040).

Park, Y., & Byrd, R. J. (2001). Hybrid text mining for finding abbreviations and their definitions. In Proceedings of the 2001 conference on empirical methods in natural language processing (pp. 126-133).

Parry, D. (2009) "Crowdsourcing" as a Means to Identify SNOMED CT Subsets–an Initial Approach.

Pennock, D. (2007). The wisdom of the ProbabilitySports crowd. Retrieved from http://blog.oddhead.com/2007/01/04/the-wisdom-of-the-probabilitysports-crowd/

Pham, D., Dimov, S., & Nguyen, C. (2005). Selection of K in K-means clustering. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 219(1), 103-119.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., & Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. ACM Transactions on Interactive Intelligent Systems (TiiS), 3(1), 3.

Qi, H., Yang, M., He, X., & Li, S. (2010). Re-examination on lam% in spam filtering. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 757-758). ACM.

Quinn, A. J., & Bederson, B. B. (2009). A taxonomy of distributed human computation. Human-Computer Interaction Lab Tech Report, University of Maryland.

Quinn, A. J., & Bederson, B. B. (2011). Human computation: a survey and taxonomy of a growing field. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1403-1412). ACM.

Radlinski, F., Bennett, P. N., Carterette, B., & Joachims, T. (2009). Redundancy, diversity and interdependent document relevance. In ACM SIGIR Forum (Vol. 43, No. 2, pp. 46-52). ACM.

Radoff, J. (2011). Game on: Energize Your Business with Social Media Games: Wiley.

Ribeiro, F., Florencio, D., & Nascimento, V. (2011). Crowdsourcing subjective image quality evaluation. In Image Processing (ICIP), 2011 18th IEEE International Conference on (pp. 3097-3100). IEEE.

Richards, C. (2003) Teach the world to twitch: An interview with Marc Prensky, CEO and founder Games2train. com. Futurelab, Dec. 2003.

Richardson, M., & Domingos, P. (2003a). Building large knowledge bases by mass collaboration. Proceedings of the 2nd international conference on Knowledge capture (pp. 129-137). ACM.

Richardson, M., & Domingos, P. (2003b). Learning with knowledge from multiple experts. ICML (Vol. 20, pp. 624-631).

Ritterfeld, U., Cody, M. J., & Vorderer, P. (2009). Serious games: Mechanisms and effects: Taylor & Francis.

Robertson, S., & Hull, D. A. (2000). The TREC-9 filtering track final report. Proceedings of the 9th text retrieval conference.

Robson, C., Kandel, S., Heer, J., & Pierce, J. (2011). Data collection by the people, for the people. Paper presented at the Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems, Vancouver, BC, Canada.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. 313-323.

Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 213-220). ACM.

Sabou, M., Bontcheva, K., Scharl, A., & Föls, M. (2013). Games with a Purpose or Mechanised Labour?: A Comparative Study. Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies (p. 19). ACM.

Salminen, J., & Harmaakorpi, V. (2011). Collective Intelligence and Practice-Based Innovation: An Idea Evaluation Method Based on Collective Intelligence. Practice-Based Innovation: Insights, Applications and Policy Implications, 213-232.

Sanchez, D., & Isern, D. (2011). Automatic extraction of acronym definitions from the Web. Applied Intelligence, 34(2), 311-327. doi: 10.1007/s10489-009-0197-4

Schwartz, A. S., & Hearst, M. (2002). A simple algorithm for identifying abbreviation definitions in biomedical text. 451-62.

Shaw, A. D., Horton, J. J., & Chen, D. L. (2011). Designing incentives for inexpert human raters. In Proceedings of the ACM 2011 conference on Computer supported cooperative work (pp. 275-284). ACM.

Simon, R., Haslhofer, B., & Jung, J. (2011). Annotations, tags & linked data. Sixth International Workshop on Digital Approaches in Cartographic Heritage, 1-8.

Simpson, J. (2004). The OED and collaborative research into the history of English. Anglia-Zeitschrift für englische Philologie, 122(2), 185-208.

Smucker, M. D., Kazai, G., & Lease, M. (2012) Overview of the TREC 2012 Crowdsourcing Track.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. Proceedings of the conference on empirical methods in natural language processing (pp. 254-263). Association for Computational Linguistics.

Sobel, D. (1995). Longitude: The true story of a lone genius who solved the greatest scientific problem of his time: Walker & Company.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. Computer, 35(3), 107-109.

Stock, W. G., & Stock, M. (2013). Handbook of Information Science. Berlin, Boston, MA: De Gruyter Saur.

Strohman, T., Metzler, D., Turtle, H., & Croft, W. B. (2005). Indri: A language model-based search engine for complex queries. Proceedings of the International Conference on Intelligent Analysis (Vol. 2, No. 6, pp. 2-6).

Su, Q., Pavlov, D., Chow, J.-H., & Baker, W. C. (2007). Internet-scale collection of human-reviewed data. Proceedings of the 16th international conference on World Wide Web (pp. 231-240). ACM.

Susi, T., Johannesson, M., & Backlund, P. (2007). Serious games–An overview. Skövde: University of Skövde (Technical Report HS-IKI-TR-07-001).

Thaler, S., Simperl, E., & Wolger, S. (2012). An experiment in comparing human-computation techniques. IEEE Internet Computing, 16(5), 0052-58.

Thompson, K. (2012). All Hands on Deck. Scientific American, 306(2), 56-59.

Tomlinson, S., Oard, D. W., Baron, J. R., & Thompson, P. (2007). Overview of the TREC 2007 Legal Track. In TREC.

Turtle, H. (1994). Natural language vs. Boolean query evaluation: a comparison of retrieval performance. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 212-220). Springer-Verlag New York, Inc.

Urbano, J., Morato, J., Marrero, M., & Martín, D. (2010). Crowdsourcing preference judgments for evaluation of music similarity tasks. ACM SIGIR Workshop on Crowdsourcing for Search Evaluation (pp. 9-16).

von Ahn, L. (2005). Human Computation. PhD Dissertation, CMU. (AAI3205378)

von Ahn, L. (2006). Games with a purpose. Computer, 39(6), 92-94.

von Ahn, L. (2013). Augmented intelligence: the Web and human intelligence. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1987).

von Ahn, L., Blum, M., Hopper, N., & Langford, J. (2003). CAPTCHA: Using hard AI problems for security. Advances in Cryptology (EUROCRYPT 2003).

von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 319-326). ACM.

von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. Communications of the ACM, 51(8), 58-67.

von Ahn, L., Ginosar, S., Kedia, M., & Blum, M. (2007). Improving image search with phetch. coustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on (Vol. 4, pp. IV-1209). IEEE.

von Ahn, L., Ginosar, S., Kedia, M., Liu, R., & Blum, M. (2006). Improving accessibility of the web with a computer game.

von Ahn, L., Kedia, M., & Blum, M. (2006). Verbosity: a game for collecting common-sense facts. Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 75-78). ACM.

von Ahn, L., Liu, R., & Blum, M. (2006). Peekaboom: a game for locating objects in images. Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 55-64). ACM.

Voorhees, E. M., & Harman, D. (1999). Overview of the eighth text retrieval conference (TREC-8). Proceedings of TREC.

Wang, A., Hoang, C. D. V., & Kan, M.-Y. (2013). Perspectives on crowdsourcing annotations for natural language processing. Language Resources and Evaluation, 1-23.

Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., & Movellan, J. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. Advances in Neural Information Processing Systems, 22, 2035-2043.

Wolf, L., Knuth, M., Osterhoff, J., & Sack, H. (2011). RISQ! Renowned Individuals Semantic Quiz: a Jeopardy like quiz game for ranking facts. Proceedings of the 7th International Conference on Semantic Systems (pp. 71-78). ACM.

Xu, J., & Huang, Y. (2007). Using SVM to extract acronyms from text. Soft Computing-A Fusion of Foundations, Methodologies and Applications, 11(4), 369-373.

Yan, T., Kumar, V., & Ganesan, D. (2010). CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones. Proceedings of the 8th international conference on Mobile systems, applications, and services (pp. 77-90). ACM.

Yang, J., Adamic, L. A., & Ackerman, M. S. (2008). Crowdsourcing and knowledge sharing: strategic user behavior on taskcn. Proceedings of the 9th ACM conference on Electronic commerce (pp. 246-255). ACM.

Yarygina, A., & Vassilieva, N. (2012). High-recall extraction of acronym-definition pairs with relevance feedback. Proceedings of the 2012 Joint EDBT/ICDT Workshops (pp. 21-28). ACM.

Yeates, S., Bainbridge, D., & Witten, I. H. (2000). Using compression to identify acronyms in text. arXiv preprint cs/0007003.

Yuen, M. C., Chen, L. J., & King, I. (2009). A survey of human computation systems. Computational Science and Engineering, 2009. CSE'09. International Conference on (Vol. 4, pp. 723-728). IEEE.

Yuen, M. C., King, I., & Leung, K. S. (2011). A survey of crowdsourcing systems. Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom) (pp. 766-773). IEEE.

Zaidan, O. F., & Callison-Burch, C. (2011). Crowdsourcing translation: professional quality from non-professionals. ACL (pp. 1220-1229).

Zhang, P., Cao, W., & Obradovic, Z. (2013). Learning by aggregating experts and filtering novices: a solution to crowdsourcing problems in bioinformatics. BMC Bioinformatics, 14(Suppl 12), S5.

Zhu, M. (2004). Recall, precision and average precision. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo. CA

Zuccon, G., Leelanupab, T., Whiting, S., Jose, J., & Azzopardi, L. (2011). Crowdsourcing Interactions. Crowdsourcing for Search and Data Mining (CSDM 2011), 35.

APPENDIX A – SCREENSHOTS FROM THE ACRONYM IDENTIFICATION AND RESOLUTION TASK

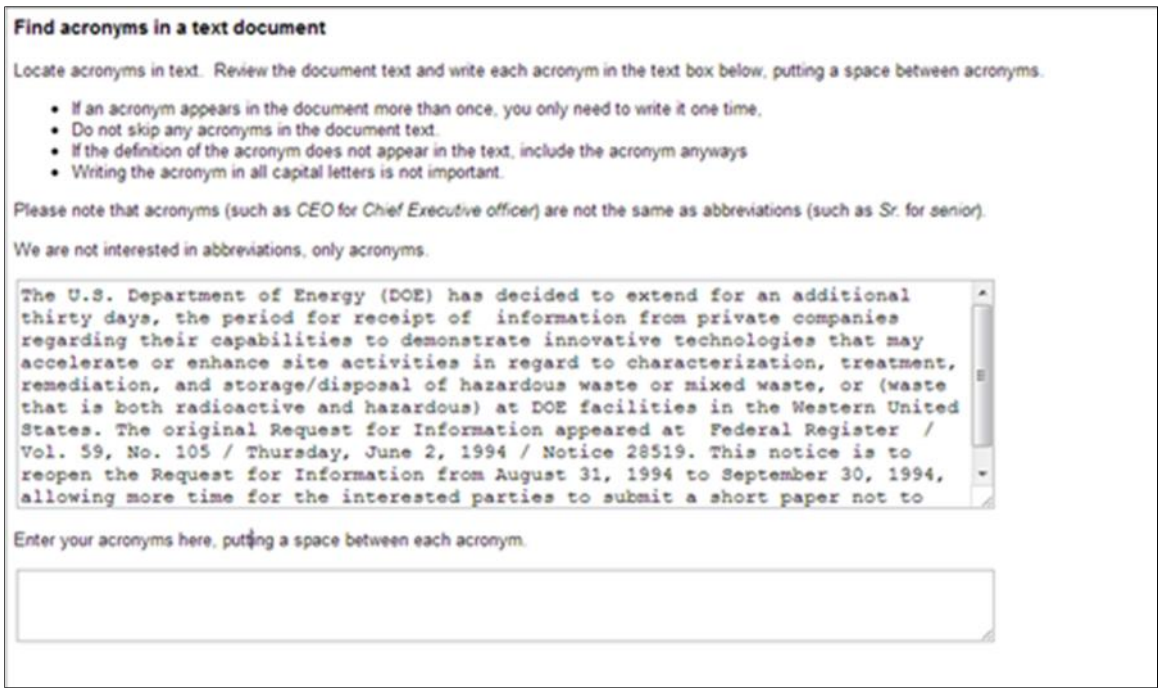

Figure A-1: Screenshot from the acronym identification phase for the non-game interface

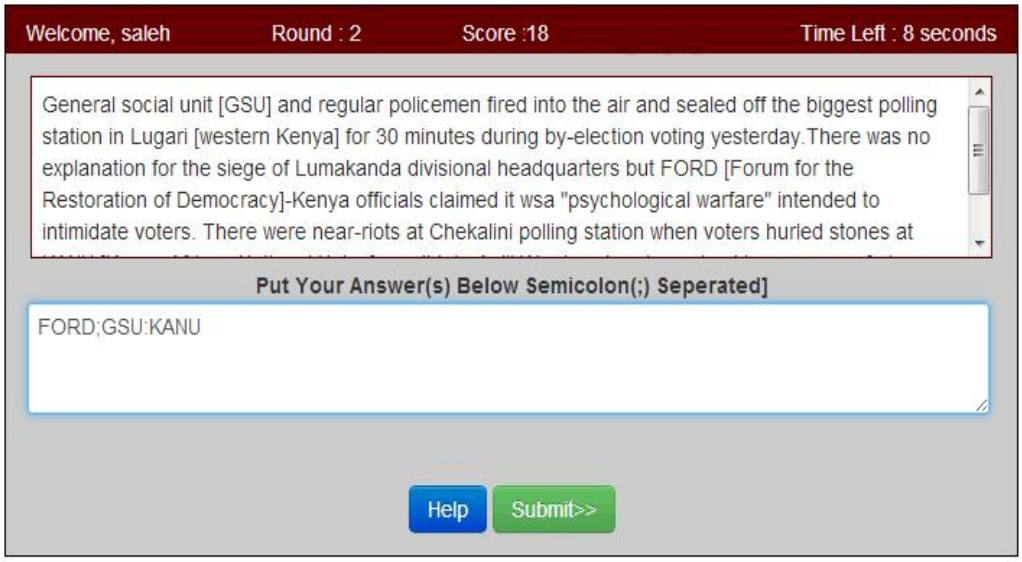

Figure A-2: Screenshot from the acronym identification phase for the game interface

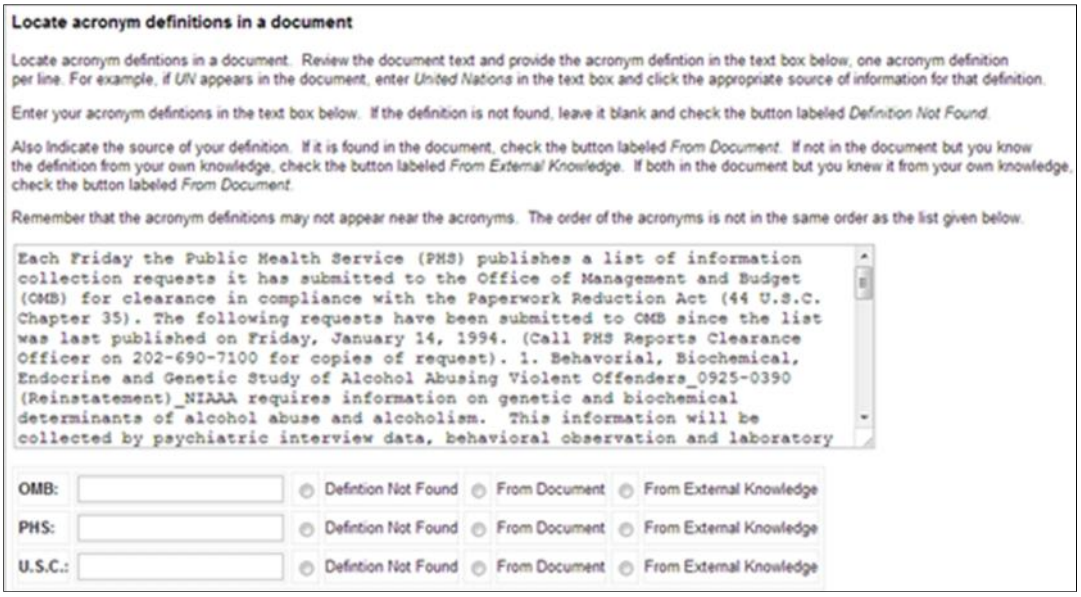

Figure A-3: Screenshot from the acronym definition resolution phase for the non-game interface

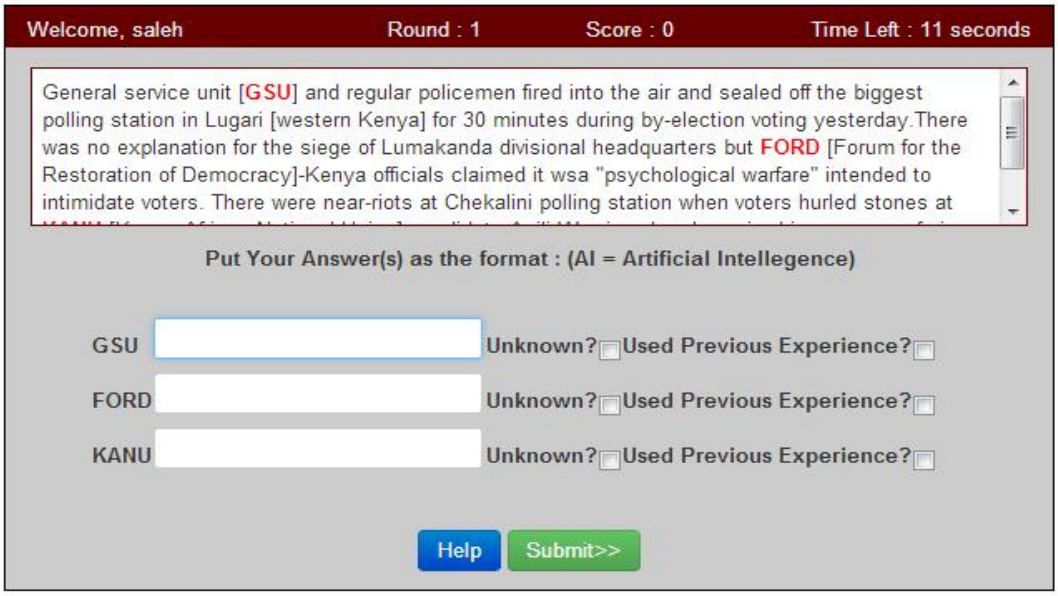

Figure A-4: Screenshot from the acronym definition resolution phase for the game interface

APPENDIX B – INSTRUCTIONS PROVIDED TO PARTICIPANTS IN THE ACRONYM IDENTIFICATION AND RESOLUTION TASK

For the acronym identification phase, they were as follows:

> Find acronyms in a text document
>
> Locate acronyms in text. Review the document and write each acronym in the text box below, putting a space between acronyms.
>
> If an acronym appears in the document more than once, you only need to write it one time.
> Do not skip any acronyms in the document text
> If the definition of the acronym does not appear, include the acronym anyways
> Writing the acronym in all capital letters is not important.
>
> Please note that acronyms (such as CEO for Chief Executive Officer) are not the same as abbreviations (such as Sr. for senior). We are not interested in abbreviations, only acronyms.
>
> Enter your acronyms here, putting a space or semicolon between each acronym.

The instructions for the acronym resolution task (Phase 2) were as follows:

> Locate acronym definitions in a document
>
> Locate acronym definitions in a document. Review the document text and provide the acronym definition in the text box below, one acronym definition per line. For example, if UN appears in the document, enter "United Nations" in the text box and clock the appropriate source of information used to obtain that definition.
>
> Enter your acronym definitions in the text box below. If the definition is not found, leave it blank if you don't know from outside knowledge and check the button labeled "definition not found"
>
> Also indicate the source of your definition. If it is found in the document, check the button labeled "from external knowledge". If both the acronym and definition are in the document but you know it from your own external knowledge, check the button labeled "from document".
>
> Remember that the acronym definitions may not appear near the acronyms. The order of the acronyms is not in the same order as the list given below.

APPENDIX C - PSEUDOCODE FOR ACRO4.PL

```
Foreach files in dir {
      Foreach sentence in file {
            Acronym = Search_acronym_function(sentence);
            Longform = Search_longform_function(Acronym, Sentence)
            If(Acronym){
                  Longform = Search_longform_function(Acronym,
            Sentence)
                  Description = Search_description_function(Acronym,
            Longform)
                  If(Description) {
                        Print Acronym - description
                  }Else{
                        Print Acronym - undefined
                  }
            }
      }
}

Search_acronym_function(sentence){
      Foreach terms in sentence{
            If(term in stop_list_terms or contains special characters
      [:;!?]){
                  Next term
            }Else{
                  If( 2 < term uppercase letters length < 10){
                        Next term
                  }Else{
                        return term as Acronym
                  }
            }
      }
}
Search_longform_function(Acronym, Sentence)
      term_window = min(|Acronym|+5, |Acronym|*2);

      # get terms from sentence with indexes

      Leftside_lf = @sentence[acronym_index – term_window …
acronym_index-1];
      Rightside_lf = @sentence[acronym_index+1 …
acronym_index+term_window];
      If( terms after acronym in parentheses () or []){
            Description = terms_in_ parentheses

            # use as description without any other checks
      }
      return (Leftside_lf, Rightside_lf)
}
Search_description_function(Acronym, Longform){
      Foreach Longform{
            @acr_letter = split Acronym by letters;
```

```
            Letter = get first element of @acr_letter;
            Foreach terms in Longform{
                  If(first letter of term  == Letter ){
                        Candidate = candidate + term;
                        If(is next element @acr_letter){
                              Letter = next element @acr_letter
                        }else{
                              NEXT_STEP;
                        }
                        Start=1
                  }
                  If(Letter in lowercase && Start == 1 ){

                        #we add some additional terms to candidate
                  string

                        Candidate = candidate + term;

                  }
            }
            NEXT_STEP:
            If( Candidate size in terms < 2)
                  Next;
            If(Candidate use less than 2 letters from Acronym)
                  Next;
            Push @Description, Candidate

            # @ - array of descriptions
      }
      Return @Description;
}
```

APPENDIX D – PERL CODE FOR ACRO4.PL

```
#!/usr/bin/perl

use File::Find;
use Lingua::EN::Sentence qw( get_sentences  );
use Cwd ('abs_path');
use Data::Dumper;
use strict ;

my $top_dir = abs_path($ARGV[0]);
my %stop_list_term = (
         'a'=>1, 'on'=>1, 'of'=>1, 'in'=>1, 'the'=>1,
         'at'=>1,'an' => 1, 'for' => 1,);
my $stop_list_regexp = join('|', keys %stop_list_term);
$stop_list_regexp = qr/\b$stop_list_regexp\b/;
my %acronyms;
sub get_acronym {
    my ($s) = @_;
    my $i = -1;
    my @acrs=();
    foreach my $w ((split(/\s+/, $s))){
        $i++;
        my $nxt;
        next if($w =~ /[:;!?]/);
        foreach my $stop (keys %stop_list_term){
            if($w =~ /^$stop$/i){
                $nxt=1;
                last;
            }
        }
        next if ($nxt);
  my $weight = $#{[($w =~ /[[:upper:]]/g)]} +1;
        $w =~ s/([(\[])?"?(\w+)("?[)\]][\., ]?)?/$2/g;
        if($weight>=2 && length $w >= 2 && length $w <= 10 &&
                $w =~ /^[a-z0-9\-\/']+$/i){
            $w=~s/^(.+)'\w$/$1/;
            my $acr = {};
            $acr->{word} = $w;
            $acr->{index} = $i;
            $acr->{weight} = $weight;
            @{$acr->{letters}} = ();
            foreach ( split(/([[:upper:]])/, $w)){
                if($_){ push @{$acr->{letters}}, lc $_; }
            }
            push @acrs, $acr;
        }
    }
    return \@acrs;
}
```

```
      sub is_start_word_as_acr {
          my ($word, $acr) = @_;
          my $fl = substr($word, 0, 1);
          if(lc $fl eq lc substr($acr,0,1)){
              return 1;
          }
          return 0;
      }
      sub check_for_lf {
          my ($s, $acr, $docno) = @_;
          my @result;
          my $term_window = ((length $acr->{word})+5 < (length $acr-
>{word})*2)?
              ((length $acr->{word})+5):(length $acr->{word})*2;
          my @words = split(/\s+/, $s);
          if(join(' ', @words[$acr->{index}+1 .. $#words]) =~
/^[(\[]"?([^)\]]+)"?[)\]]/){
              push @{$acronyms{$docno}->{$acr->{word}}}, $1;
          }
          my @words = map {s/([(\[])?(\w+)([)\]])?/$2/; $_} @words;
          my $left_boundary = ($acr->{index}-$term_window >=
0)?($acr->{index}-$term_window):0;
          my $right_boundary = ($acr->{index}+$term_window <=
$#words)?($acr->{index}+$term_window):$#words;
          my @ls_lf = @words[$left_boundary .. $acr->{index}-1];
          my @rs_lf = @words[$acr->{index}+1 .. $right_boundary];
          foreach(@ls_lf){
              if(is_start_word_as_acr($_, $acr->{word})){
                  push @result, [@ls_lf];
                  last;
              }
          }
          foreach(@rs_lf){
              if(is_start_word_as_acr($_, $acr->{word})){
                  push @result, [@rs_lf];
                  last;
              }
          }
          return \@result;
      }
      sub some_checks_for_description {
          my ($candidate, $acr, my $fl_acr) = @_;
          if( $candidate =~ /[:;?!,]/ or $candidate !~ /[[:lower:]]/
                  or $candidate =~ $acr->{word} or $candidate =~
/^\s+$/){
              return 0;
          }
          if(length $acr->{word} - $fl_acr >= 2 ){
              return 0;
          }
      ## print"=xx ",( grep {!/$stop_list_regexp$/} split(/ /,
$candidate)),"\n";
```

```
          if( scalar ( grep {!/^$stop_list_regexp$/} split(/ /,
$candidate)) < 2){
              return 0;
          }
          my %tmp = ();
          foreach (split(/ /, $candidate)){
              if(exists $tmp{$_}){
                  return 0;
              }else{
                  $tmp{$_}=1;
              }
          }
          return 1;
      }
      sub get_description {
          my ($acr, $lf) = @_;
          my @candidates = ();
          foreach my $longform (@$lf){
              my @fl_acr;
              my %w_for_skip = ();
              my ($acr_l, $_exit, $candidate, $last_term) = ({}, 0,
'', '');
              my @tmp_candidates = ();
              while(!$_exit){
                  my ($skip, $start, $check) = (0,0,0);
                  $candidate = '';
                  @fl_acr = map {($_ =~ /[A-Z]/)?({lc
$_=>1}):({$_=>0})} split(//, $acr->{word});
                  my $acr_l = shift @fl_acr;
                  while((values %$acr_l)[0] == 0 ){
                      $acr_l = shift @fl_acr;
                  }
                  foreach my $lf_term (@$longform){
      # # print$acr->{word} , " $lf_term \n";
                      next if (exists $w_for_skip{$lf_term});
                      if(! exists $acr_l->{(lc substr($lf_term, 0,
1))} && $start &&
                              $skip == 0 && $#fl_acr >= 0
                              && !exists $stop_list_term{$lf_term}
                        )
                      {
      # # print"candidate = $candidate\n";
                          $w_for_skip{$last_term}=1;
                          if(some_checks_for_description($candidate,
$acr, $#fl_acr)){
                              push @tmp_candidates, $candidate;
      # # printDumper \@candidates;
                          }
                          $check = 1;
                          last;
                      }
                      $last_term = $lf_term;
```

```
                    if(exists $acr_l->{(lc substr($lf_term, 0,
1))} &&
                            $acr_l->{(lc substr($lf_term, 0, 1))}
== 1 ){
                        if($skip){$skip=0;}
                        unless($start){$start=1;}
                       # print"add $lf_term \n";
                        $candidate .= "$lf_term ";
                       # print$candidate,"\n";
                        $acr_l = shift @fl_acr;
                        next;
                    }
                    if((exists $acr_l->{(lc substr($lf_term, 0,
1))} && $start &&
                                $acr_l->{(lc substr($lf_term, 0,
1))} == 0) || $skip == 1){
                        unless($skip){$skip = 1;}
                       # print"add1 $lf_term \n";
                        $candidate .= "$lf_term ";
                    }
                    if(exists $stop_list_term{$lf_term} && $start
&& $#fl_acr >= 0){
                       # print"add2 $lf_term \n";
                        $candidate .= "$lf_term ";
                    }
                }
                if($check == 0){
                    $_exit = 1;
                }

            }
            if(some_checks_for_description($candidate, $acr,
$#fl_acr)){
                push @candidates, $candidate;
     ## printDumper \@candidates;
            }elsif($#tmp_candidates >= 0){
                push @candidates, pop @tmp_candidates;
            }
        }
        return \@candidates;
    }

    sub acronym {
        my ($docno, $lines) = @_;
        $lines=~s/\n\n/\.\n/mg;
        my $sentences = get_sentences($lines);
        foreach my $s (@$sentences){
            if($s =~ /--/){
                $s = (split(/--/, $s))[1];
            }
            next if $#{[($s =~ /[[:lower:]]/g)]} == -1;
            my $acrs = get_acronym($s);
```

```
            foreach my $acr  (@$acrs){
                my $lf_for_check = check_for_lf($s, $acr, $docno);
                my $descr;
                if($acr->{word} && $#{$lf_for_check}>-1){
                    $descr = get_description($acr, $lf_for_check);
                }else{
                    push @{$acronyms{$docno}->{$acr->{word}}},
'undefined';
                    next;
                }
                if($#{$descr} >= 0){
                    foreach(@$descr){
                        push @{$acronyms{$docno}->{$acr->{word}}},
$_;
                    }
                }else{
                    push @{$acronyms{$docno}->{$acr->{word}}},
'undefined';
                }
            }
        }
    }
    sub wanted {
        my $fn = $File::Find::name;
        my ($doc, $docno, $lines);
        if(open(F, $fn)){
            while(my $str = <F>){
                if($str =~ m#<DOCNO>\s?(\S+)\s?</DOCNO>#){
                    $docno = $1;
                }
                if($str =~ /<TEXT>/){ $doc = 1; }
                if($str =~ /(<\/TEXT>|<\/DOC>)/){
                    $doc = 0;
                    if($docno){
                        if($lines=~/\w+/){
                            acronym($docno,$lines);
                        }
                        $docno = 0;
                    }
                    $lines='';
                }
                if($doc){
                    if($str=~/<[^>]+>.+<[^>]+>/){
                        next;
                    }
                    $lines .= $str;
                }
            }
            close F;
        }

    }
```

```
find(\&wanted, $top_dir);

foreach my $docno (keys %acronyms){
    foreach my $acr (keys %{$acronyms{$docno}}){
        my $defined = 0;
        foreach my $descr (@{$acronyms{$docno}->{$acr}}){
            if($descr ne 'undefined'){ $defined = 1; }
        }
        my %tmp = ();
        unless($defined){
            print"$docno $acr - undefined\n";
            next;
        }
        foreach my $descr (@{$acronyms{$docno}->{$acr}}){
            $descr=~s/\s$//g;
            if($defined && $descr ne 'undefined'){
                unless(exists $tmp{$descr}){
                     print"$docno $acr - $descr\n";
                    $tmp{$descr}=1;
                }
            }
        }
    }
}
```

APPENDIX E – LIST OF STOPWORDS USED

able
about
above
abroad
according
accordingly
across
actually
adj
after
afterwards
again
against
ago
ahead
ain't
all
allow
allows
almost
alone
along
alongside
already
also
although
always
am
amid
amidst
among
amongst
an
and
another
any
anybody
anyhow
anyone
anything
anyway
anyways
anywhere
apart
appear
appreciate
appropriate
are
aren't
around
as
a's
aside
ask
asking
associated
at
available
away
awfully
back
backward
backwards
be
became
because
become
becomes
becoming
been
before
beforehand
begin
behind
being

believe
below
beside
besides
best
better
between
beyond
both
brief
but
by
came
can
cannot
cant
can't
caption
cause
causes
certain
certainly
changes
clearly
c'mon
co
co.
com
come
comes
concerning
consequently
consider
considering
contain
containing
contains
corresponding
could
couldn't
course
c's
currently
dare
daren't
definitely
described
despite
did
didn't
different
directly
do
does
doesn't
doing
done
don't
down
downwards
during
each
edu
eg
eight
eighty
either
else
elsewhere
end
ending
enough
entirely
especially
et
etc
even
ever
evermore
every
everybody

everyone
everything
everywhere
ex
exactly
example
except
fairly
far
farther
few
fewer
fifth
first
five
followed
following
follows
for
forever
former
formerly
forth
forward
found
four
from
further
furthermore
get
gets
getting
given
gives
go
goes
going
gone
got
gotten
greetings
had
hadn't
half
happens
hardly
has
hasn't
have
haven't
having
he
he'd
he'll
hello
help
hence
her
here
hereafter
hereby
herein
here's
hereupon
hers
herself
he's
hi
him
himself
his
hither
hopefully
how
howbeit
however
hundred
i'd
ie
if
ignored

i'll
i'm
immediate
in
inasmuch
inc
inc.
indeed
indicate
indicated
indicates
inner
inside
insofar
instead
into
inward
is
isn't
it
it'd
it'll
its
it's
itself
i've
just
k
keep
keeps
kept
know
known
knows
last
lately
later
latter
latterly
least
less
lest
let
let's
like
liked
likely
likewise
little
look
looking
looks
low
lower
ltd
made
mainly
make
makes
many
may
maybe
mayn't
me
mean
meantime
meanwhile
merely
might
mightn't
mine
minus
miss
more
moreover
most
mostly
mr
mrs
much
must

mustn't
my
myself
name
namely
nd
near
nearly
necessary
need
needn't
needs
neither
never
neverf
neverless
nevertheless
new
next
nine
ninety
no
nobody
non
none
nonetheless
noone
no-one
nor
normally
not
nothing
notwithstanding
novel
now
nowhere
obviously
of
off
often
oh
ok
okay
old
on
once
one
ones
one's
only
onto
opposite
or
other
others
otherwise
ought
oughtn't
our
ours
ourselves
out
outside
over
overall
own
particular
particularly
past
per
perhaps
placed
please
plus
possible
presumably
probably
provided
provides
que
quite

qv
rather
rd
re
really
reasonably
recent
recently
regarding
regardless
regards
relatively
respectively
right
round
said
same
saw
say
saying
says
second
secondly
see
seeing
seem
seemed
seeming
seems
seen
self
selves
sensible
sent
serious
seriously
seven
several
shall
shan't
she
she'd
she'll
she's
should
shouldn't
since
six
so
some
somebody
someday
somehow
someone
something
sometime
sometimes
somewhat
somewhere
soon
sorry
specified
specify
specifying
still
sub
such
sup
sure
take
taken
taking
tell
tends
th
than
thank
thanks
thanx
that
that'll

thats
that's
that've
the
their
theirs
them
themselves
then
thence
there
thereafter
thereby
there'd
therefore
therein
there'll
there're
theres
there's
thereupon
there've
these
they
they'd
they'll
they're
they've
thing
things
think
third
thirty
this
thorough
thoroughly
those
though
three
through
throughout
thru
thus
till
to
together
too
took
toward
towards
tried
tries
truly
try
trying
t's
twice
two
un
under
underneath
undoing
unfortunately
unless
unlike
unlikely
until
unto
up
upon
upwards
us
use
used
useful
uses
using
usually
v
value
various

versus
very
via
viz
vs
want
wants
was
wasn't
way
we
we'd
welcome
well
we'll
went
were
we're
weren't
we've
what
whatever
what'll
what's
what've
when
whence
whenever
where
whereafter
whereas
whereby
wherein
where's
whereupon
wherever
whether
which
whichever
while
whilst
whither
who
who'd
whoever
whole
who'll
whom
whomever
who's
whose
why
will
willing
wish
with
within
without
wonder
won't
would
wouldn't
yes
yet
you
you'd
you'll
your
you're
yours
yourself
yourselves
you've

APPENDIX F - DOCUMENTS EVALUATED BY AT LEAST 25% OF THE CROWD

Table F-1: News collection documents the crowd says are relevant, but TREC Assessors said are non-relevant:

Topic ID	Document ID
421	FR940511-1-00077
421	FT922-14197
421	FT943-4815
436	FT932-5424
436	LA020189-0055
436	LA032790-0024
436	LA071390-0080
436	LA092090-0244
436	LA100490-0053
436	LA112989-0069

Table F-278: News collection documents the crowd says are not non-relevant, but TREC assessors said are relevant:

Topic ID	Document ID
421	FR940919-0-00118
421	FT931-14481
421	FT932-11383
421	FT934-16300

Table F-3: News collection documents the crowd says are relevant, but TREC assessors did not evaluate:

Topic ID	Document ID
403	FR940303-1-00044
421	FR940728-0-00054
421	FT943-14326
421	LA090990-0137
436	FBIS3-24177
436	FBIS3-3537
436	FBIS3-41760
436	FBIS3-60668
436	LA011289-0070

Table F-4: OHSUMED collection documents the crowd says are relevant, but TREC assessors said are non-relevant:

Topic ID	Document ID
13	89009548
13	89255639
13	90354721

Table F-5: OHSUMED collection documents the crowd says are not non-relevant, but TREC assessors said are relevant:

Topic ID	Document ID
12	90208451

APPENDIX G – INSTRUCTION SCREENS USED FOR SEEK-O-RAMA AND SEEKGAME

G.1 Seek-o-rama (Non-game interface)

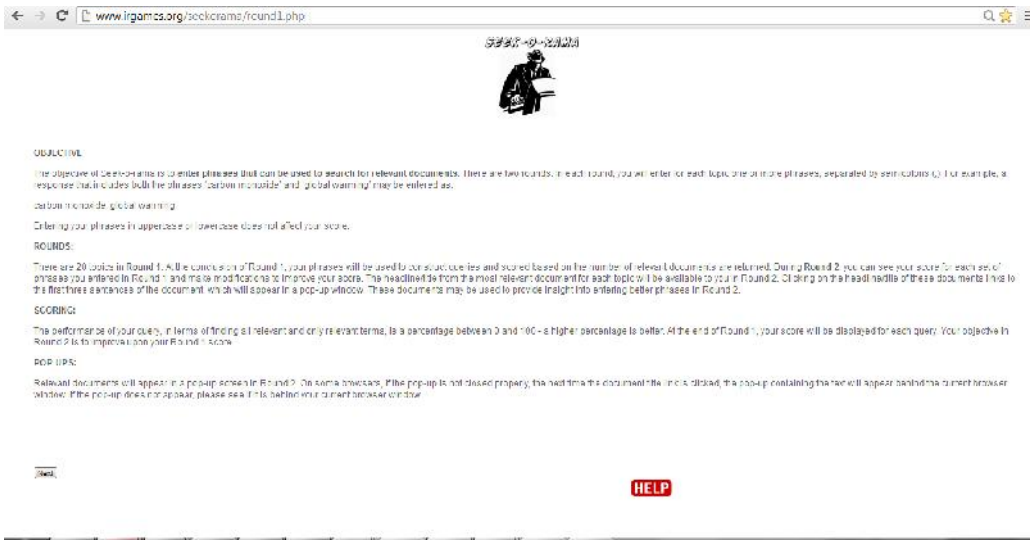

Figure G-1: Instruction screen for the News collection using Seek-o-Rama

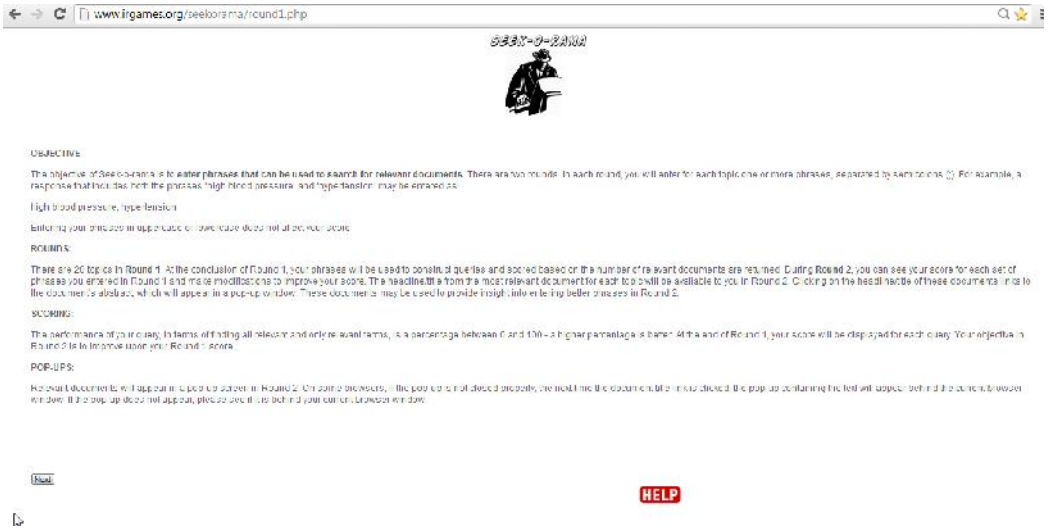

Figure G-2: Instruction screen for the OHSUMED collection using Seek-o-Rama

G.2 Seek game (Game interface)

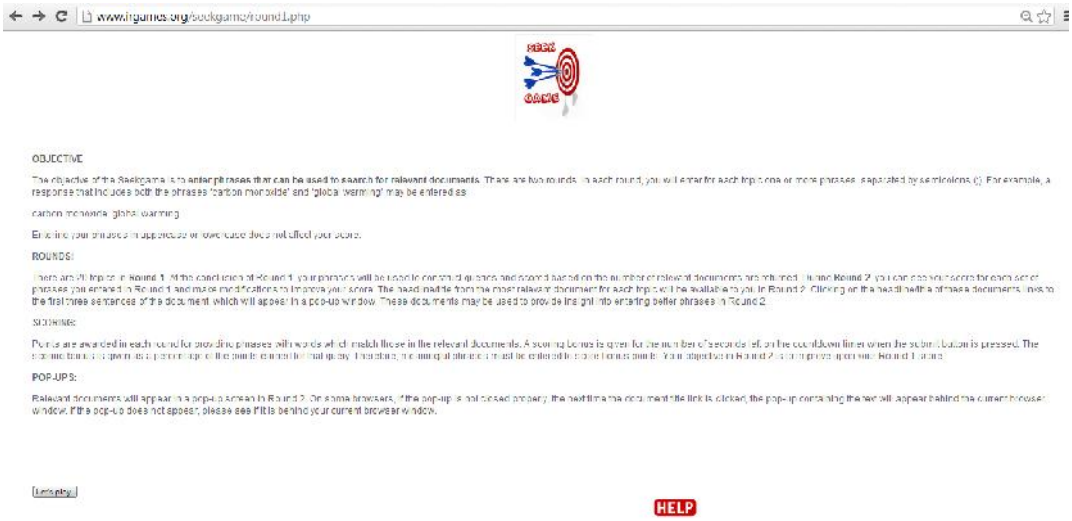

Figure G-3: Instructions for the Seekgame in the News collection

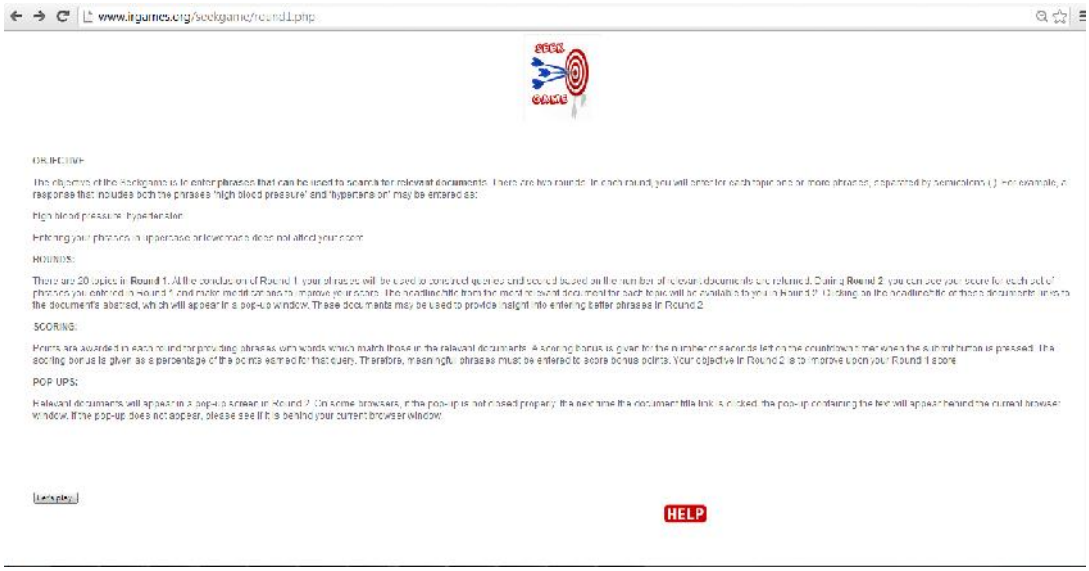

Figure G-4: Instructions for the Seekgame in the OHSUMED collection