SIMULATING QUESTION-BASED VISUAL SCANNING FOR NON-VISUAL READERS

by

Debra Yarrington

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science.

Summer 2019

SIMULATING QUESTION-BASED VISUAL SCANNING FOR NON-VISUAL READERS

by

Debra Yarrington

Approved: _____
Kathleen F. McCoy, Ph.D.
Chair of the Department of Computer and Information Sciences

Approved: _____
Levi T. Thompson, Ph.D.
Dean of the College of Engineering

Approved: _____
Douglas J. Doren, Ph.D.
Interim Vice Provost for Graduate and Professional Education and Dean of the Graduate College

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Kathleen F. McCoy, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Sandra Carberry, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Vijay Shanker, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Beth Mineo, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

This dissertation would not have been possible without the help and guidance of some very special and notably patient people in my life.

First, I must thank my advisor, Dr. Kathleen F. McCoy. Throughout the many years I've known and worked with her she has not only been a mentor and advisor, but also a role model of a strong, intelligent, compassionate woman, a close friend, and a source of endless amusement. I've always relied on her wisdom and guidance with my dissertation, my career, and my life in general, and I feel incredibly fortunate to have her support and friendship.

I would like to thank my committee, Dr. Beth Mineo, Dr. Vijay Shankar, and Dr. Sandee Carberry for their time, insightful comments and their suggestions for the direction my dissertation should take. I must also thank them for their patience.

I would like to thank Dr. Timothy Bunnell, who was one of the first people in my career to believe in me and support me. It was Tim who first planted the idea of getting a Ph.D., and it was his guidance that led me to grow and move on in my career. I admire his dedication and his fairness, and I am grateful for his support.

I would like to thank my consultants, Elizabeth Bottner, Tanya Servic, and Ken Rolph. I first tutored Liz Bottner 14 years ago. She is largely responsible for my being made aware of the issues that people who are blind encounter. It was the many hours we spent on homework assignments that led me to believe there must be a way to improve the learning experience for people who are nonvisual readers. Liz has long ago graduated, yet whenever I need it she is always willing to help. She gladly agreed

work on the interface with me and gave invaluable feedback. I am very grateful for that.

I would like to thank Tanya Servic. Tanya is someone who is always friendly and cheerful and always willing to help, in spite of all she has going on in her life. Tanya was another student who inspired this dissertation years ago, and has provided invaluable feedback on the low-vision experience throughout this dissertation.

I can't thank Ken Rolph enough for his unfailing help in giving me guidance with the system and the user interface. Ken was amazingly patient and always willing to look at the latest version of the interface. His feedback was precise, clear, and overall invaluable. This dissertation would not have been possible without his input. I am very grateful that I was able to find him and work with him.

I would like to thank my friend Kathy Samworth, who has been through so much with me. She has heard more about the ups and downs of getting a PhD than anyone should have to endure, yet was always supportive and encouraging.

There's really no way to properly express my gratitude to my parents. My parents managed to simultaneously convey to me that I was the most special person in the world and that nothing was going to be handed to me in life. Being incredibly supportive without ever coddling is a fine line to hold, but they did, always. I am so very grateful for the foundation they gave me. I am also so very grateful for their unending love. How lucky am I.

Finally I would like to gratefully thank my husband Steve. I am incredibly fortunate to have so many remarkable people in my life, and I am even more fortunate that one of those people is my husband. Steve, you're my safe harbor, my sounding board, and my kick in the butt. I adore you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This dissertation describes the creation of a system for locating information in a text document that is relevant to a complex question. While the system can be used by anyone to efficiently identify text areas related to a question within a large amount of less relevant text, it was specifically designed for non-visual readers, notably people who are blind and low-vision. Visual readers often quickly scan through documents to locate relevant information, yet non-visual readers have few options for intelligently scanning documents for information relevant to complex questions. This can reduce efficiency in answering homework questions, in obtaining relevant information in work documents, in learning new information for enjoyment, and even in retrieving information from a previous pass through a text document. Thus the purpose of this dissertation is to develop a system that allows non-visual readers to gather information related to complex questions as quickly and efficiently as their visually-scanning counterparts.

This dissertation is subdivided into 3 parts:

1. Determining how visual readers scan through documents when answering complex questions;

2. Developing and implementing a method that replicates the identification and location of relevant text within a document similar to what visual readers identified as determined by part 1; and

3. Developing a user interface to allow users to move through a document so that they garner all or most information garnered by visual scanners, including not only question-relevant information, but also topological information and information surrounding the relevant information.

Chapter 1

INTRODUCTION

This dissertation describes a system that allows users to scan through a document in response to a complex question and quickly and efficiently locate and acquire information related to the question. It was designed to allow non-visual readers to gather the same or similar information that a visual reader would gain when scanning through the document to locate the answer to a question. While this system can be used by anyone to improve efficiency and accuracy in quickly identifying all question-relevant textual information in text documents, the long-term goal of this research is to improve the educational and professional opportunities of people who use assisted reading technologies such as a screen reader or a screen magnifier (henceforth referred to as non-visual readers). While this group includes, but is not limited to, those with visual impairments, those who tire easily visually reading, and those who have a learning disability, such as dyslexia, that affect their ability to visually read, the specific groups focused on in this dissertation are readers who are blind and low vision. The scanning system works in conjunction with existing accessibility software (such as screen readers and screen magnifiers). The goal is to allow users of these technologies to scan a document to answer a complex question while simultaneously acquiring information about the question's topic and the overall document content similar to the information someone who visually scans a document would acquire in a comparable amount of time.

A complex question is defined as one in which a simple, fact-based answer is insufficient. For instance, “When was Lincoln born?” is a simple question with a straightforward answer. An example of a complex question, on the other hand, would be “In Pride and Prejudice, how does the title relate to the characters in the book?” Complex questions usually require a good deal more comprehensive information in order to be answered. In the process of answering complex questions, readers usually need to access information within a document that cannot be found by a simple pattern matching. Readers often need to access information in more than one place within a document, and often need to read text surrounding the most relevant text in order to comprehensively understand the answer to the question. This dissertation was designed to allow users, especially non-visual readers, to access all information relevant to answering a complex question, including surrounding information and information in different locations within the document.

1.1 Motivation

The work in this dissertation was motivated by experiences I had while working with students who used assisted reading technologies provided by the ADA Office at the University of Delaware. One of the greatest difficulties these students faced was the enormous amount of time they had to spend in finding places in documents where their questions could be answered. The problem was encountered whether a document was in Braille, magnified by a screen magnifier (software that increases the size of text on a screen), or read with a screenreader (software that reads electronic text aloud). For instance, I once spent almost 8 hours helping a student who is blind take a final exam that was scheduled to take 2 hours. The exam in ink-print was 10 pages, but in Braille printed out on over 40 pages. Many of the questions

referred to paragraphs and text elsewhere in the document. Finding the text that the paragraphs referred to was very tedious and time consuming, and significantly slowed the progress of the exam, even with me there to help. The student could not easily and quickly scan through the document to find the text she needed.

The inability to quickly scan through documents for relevant text made schoolwork significantly more difficult for the students I worked with. My experience reflects the general tendency for students using alternatives to visual reading to quit school earlier than their visual-reading counterparts. Indeed, of the estimated 10 million blind and visually impaired individuals in the United States today (American Foundation for the Blind, 2007), it is well documented that these populations fall behind in education, which often results in unemployment or underemployment and subsequently affects socio-economic status as well (National Center for Policy Research for Women and Families, 2004; US Department of Labor, 2007; Wagner & Valdes, 1995).

As I found when assisting my student taking her test, reading Braille doesn't put students on par with the general population because it doesn't easily permit scanning documents. In addition, many students object to Brailled documents because their sheer bulkiness makes them difficult to transport and navigate[1]. Indeed, today, fewer than 10% of legally blind people in America have learned Braille and only 10% of blind children are learning it according to the National Federation for the Blind. Many students today (including the students I worked with) prefer to use a screenreader (e.g., JAWS, from Freedom Scientific, Window Eyes from GW Micro, HAL Screen Reader from Dolphin Computer Access). Students often configure the

[1] The size of Braille is approximately 24 point type.

screenreader to speak at quite fast rates. While the normal rate of speaking is about 180 words per minute, the students were quite comfortable listening to speech sped up to 400 and 500 words per minute. However, even at rates of 500 words per minute, it takes these students significantly longer to get through a document than it does a person who is visually scanning through the document, especially when scanning for specific information within the document.

JAWS, the most commonly used screenreader, does offer some rudimentary scanning options. The screen reader lets one navigate through a document by reading the first line or sentence of each paragraph. JAWS also allows one to create text rules used to locate passages in a document that contain a particular word or phrase and then returns a set of links to the lines, sentences, or paragraphs (depending on the setting you choose) containing that word or phrase. JAWS even allows for the use of regular expressions to more precisely control the search of a document. However, these options are not as useful as they may seem because passages relevant to answering questions often contain semantically related words rather than exact repetitions of words in the question. JAWS 11.0 includes a tool for creating a word index, or list of words that appear in a document. The words are ordered by the number of times they occur within the document, and are linked to their location in the document. It also allows individuals to create a list of sentences that contain a particular word, with each sentence being a link to its location within the document. However, none of these options use any intelligent reasoning mechanisms to figure out where within a document to jump to next. This leaves people who rely on these technologies at a serious disadvantage. Consider the following tasks: after reading a document, finding the paragraph that discussed the author's description of writing her first novel so it

could be reread; finding an area in a text that described the migratory patterns of the Beluga Whale. None of these things could be found with the simple scanning technologies offered in current screen reader technology when trying to answer homework questions from a text. Students using screenreaders have little choice other than spending enormous amounts of time reading the entire document from the beginning until they hit upon text relevant to the question. Even after the answering text was found, when students needed to hear the text again, they often had to start re-listening from the beginning. This put them at an incredible disadvantage compared to their classmates who could accomplish the same task in a fraction of the time.

> "I usually just end up reading everything; I don't have the benefit of just skimming the paper for the answers or only reading half of it. Unfortunately, nothing beats having an actual hardcopy in front of me, although even skimming with Braille can be somewhat tedious. You'd think there'd be some way around that, or some way to make it easier..."
>
> *Liz Bottner,*
> *University of Delaware 2008 Graduate*

1.2 Scanning System

The system described in this dissertation takes a single text document and a complex question or questions, and returns places in the document that are likely to be most connected to the question. The system attempts to determine which information is most connected to the question using algorithms developed to replicate what visual readers focus on when scanning through documents. In particular, algorithms were created that attempted to identify the same information that sighted scanning individuals spent the most time focused on it when scanning for the question's answer and thus appear to have found to be of most interest. These algorithms have been verified to work across a variety of topics by comparing their output with eye tracking

data collected from visually-reading individuals. The system allows users to quickly switch between different modes of navigation within the document (e.g., to step through the document to the most important sentences first or to step through important sentences in the order they appear in the document) and to easily read information located physically near important sentences. By allowing users to access the information in different modes, and allowing the user to quickly switch between the different modes, the user is able to access this information in ways that provide not just material directly related to the question, but also other information about the document contents (e.g., about the topology of the document, what the document is "about").

The aim is that users of this system will have an experience similar to a visually scanning experience and that users will get information similar to the information visual scanners get when scanning the document for an answer in a manner that is as efficient as visual scanners. The program was designed to be useful for individuals who are blind and visually impaired with the expectation that it may be modified for use by those with dyslexia and other learning abilities that make visual reading difficult, and for those who fatigue easily (for instance, those with cerebral palsy).

1.3 Research Objectives

The fundamental objectives of this dissertation were:

4. To systematically identify the information individuals gather when visually scanning through a document in order to answer a question. For this the Eye Tracker System from Tobii Technology was used to gather information about the text individuals focus on when scanning through documents to answer a question. The level of analysis I have chosen here is the paragraph level;

5. To develop NLP methods that enabled the system to identify the importance of various paragraphs in answering a question;

6. To evaluate the methods in objective 2 by comparing results with eye tracking data from objective 1.

7. To develop a software system that incorporated the NLP analysis and output importance measures for paragraphs.

8. To develop a user interface that effectively used the results of the developed software system; and

9. To ensure usability and usefulness of the system through studies and feedback from potential end users.

It is important to note that the goal of the system was not to return to the user the answer to the question, but to create a system that allowed users to have an experience similar to scanning through the document when answering a question. With the system, the user should be able to not only locate relevant text, but also to learn where within the document the relevant text was located, to be able to navigate easily to text surrounding the text identified as relevant to gather more information, and to be able to learn where a large number of identified-as-relevant text was located so users would know that focusing on that area would likely be beneficial.

Producing the system required the attainment of three major goals:

(1) Achieving an understanding of how visual scanners processed a document when scanning and what information in the document they paid more attention to when scanning in response to a complex question

(2) Developing Natural Language Processing (NLP) techniques to automatically identify the text in documents visual readers focused on as determined in step 1

(3) Developing a user interface to be used in conjunction with screen reading software to deliver the visual scanning experience, including not only identifying the information focused on by visual scanners, but also obtaining an overall topology of the document and where relevant information was located topologically within the document.

The research methodology used in this dissertation was one of user centered design and iterative refinement. Part (1) made use of extensive user studies with eye-tracking technology to identify how visual readers scanned, and to quantify measurable parameters to be reproduced by the system. The measurement of the success of the NLP techniques developed in Part (2) was against these parameters – the system had to identify as important those areas in the text that are important to visual scanners. Finally, in the user interface design and development in Part (3) user centered design (Norman, 1988) and participatory design with eventual end users of the technologies were used.

This system can be broken down into individual components, e.g., the user interface, the identification of what should be considered relevant, and the methodology for locating that relevant text within a document. While this dissertation was designed with each of these components focusing on a particular problem or group of potential end users, each component should be modifiable to adapt or expand the system to different needs and different end users. For instance, while the NLP component of the system was designed a number of years ago to mimic our findings of where visual scanners spent time in a document, as newer NLP techniques emerge (and have emerged), it may be beneficial to the system to adapt the NLP component to include these new techniques to more faithfully identify relevant text. Equally, while

the user interface was designed for individuals who are blind and low vision, it is hoped that that component could be modified to better suit the needs of individuals who are dyslexic, individuals who fatigue easily while reading, and even visual readers who may wish to find relevant text in a document efficiently. Thus the system is a prototype for a system that may hopefully have broad applications beyond its original design.

Chapter 2

UNDERSTANDING VISUAL SCANNING

My goal in developing intelligent scanning software was to provide a scanning experience similar to that of a visual scanner. Intuitively, when a visual reader scans a document to find an answer to a question, s/he does not just come away with the answer to the question: s/he gains knowledge about information in the document related to the question, knowledge of the document itself (its topology), and knowledge of the document domain (significantly more than the answer to the question). Visual scanning has been studied to some extent in the psychological literature (Raynor, 1998, Dyson et al., 2000, Wilkinson et al., 2006), but it is difficult to see how to apply these findings to the task of conveying information gleaned while scanning.

For my purposes, I was interested in what text readers focused on when scanning in connection to a question. While many systems exist that focus on answering simple, fact-based questions, my interests differed from this. I was interested in what scanners focused on when answering more complex questions in which the answer couldn't be found using pattern matching and in which the answer required at least a few sentences, not necessarily contiguous within a document. From an NLP standpoint, locating longer answers in relation to a question that a) may require gathering information from more than one place in a document; and b) may or may not have words or word sequences in common with the question posed an interesting and difficult problem. The problem became making semantic connections

within any domain that were more loosely associated than the synonyms, hypernyms, hyponyms, etc. provided by WordNet (Felbaum, 1998). From my experience, the questions that students had the most difficulty with were more complex in nature. Thus I wanted to find out what visual scanners focused on when scanning for the answer to complex questions. I wanted to know whether visual scanners were able to locate text in documents relevant to complex questions and, if so, what connections the visual scanners were making in terms of the text they chose to focus on. My approach was to learn what factors were important when scanning a document by learning where visual readers focused during the scanning process. In order to acquire this knowledge, I conducted a series of experiments using eye-tracking technologies.

2.1 Data Collection/Task Description

To identify how visual readers scan documents to answer questions, I collected 12 questions obtained from students' homework assignments, along with the documents from which the answers could be obtained.

The documents used for the experiment were text documents with very few other physical markers. None of the documents had images, figures or graphs. Two of the documents included a numbered list, and two of the documents had head sections (three or fewer).

The questions chosen were on a wide variety of topics and were more complex in nature than simple, fact based questions. An example of a typical question is, "According to Piaget, what techniques do children use to adjust to their environment as they grow?" Questions used can be found in Appendix A. Ten of the documents from which the answers could be obtained were two pages in length, one document was eight pages in length, and one document was nine pages long. In each case, the

answer to the question could be found within a single paragraph in the document, although relevant information was often found throughout the document and the answer itself was not found verbatim in the document. In all cases, the answer was to be constructed from the particular information contained in the document.

2.1.1 Task Description

Forty-three visual reading subjects scanned for the answer to between 6 – 12 questions. The subjects sat in front of a computer screen to which the Eye Tracker 1750 by Tobii Technologies was installed. The questions and accompanying documents were displayed on the computer screen and, after being calibrated, subjects were tracked as they scanned for the answer. For the two-page documents, the question appeared at the top of the first page. For the longer documents, the question appeared at the top of each page. When done scanning each document, subjects were asked to select a best answer in multiple choice form (to give them a reason to take the scanning task seriously).

2.1.2 Results

Results showed that subjects were reliably able to correctly answer the multiple choice question after scanning the document. Of the 510 questions, 423 (about 86%) were answered correctly. The two longer questions were the least likely to be answered correctly (one had 10 correct answers of 21 total answers, and the other had 10 incorrect answers and only one correct answer). On the other hand, five of the two-page questions were either always answered correctly or were answered incorrectly only once.

Clearly for the shorter documents, subjects were able to successfully answer the question. With that established, I was interested in analyzing the eye tracking data to see if there was a connection between where subjects spent the most time in the document and the question. If there was an understandable connection, the goal then became to automatically replicate those connections and thus automatically locate places in the text where subjects were most likely to spend the most time.

The Tobii Eye Tracking System records the track of a subject's eye gaze as the subject reads through a document on the computer screen and keeps track of the path of the eye gaze and the length of time the subject's gaze stayed at any particular spot. It produces a video of the eye's movements through the document, a gaze plot showing the ordered plot points of the gaze within the document (see Figure 1), and a

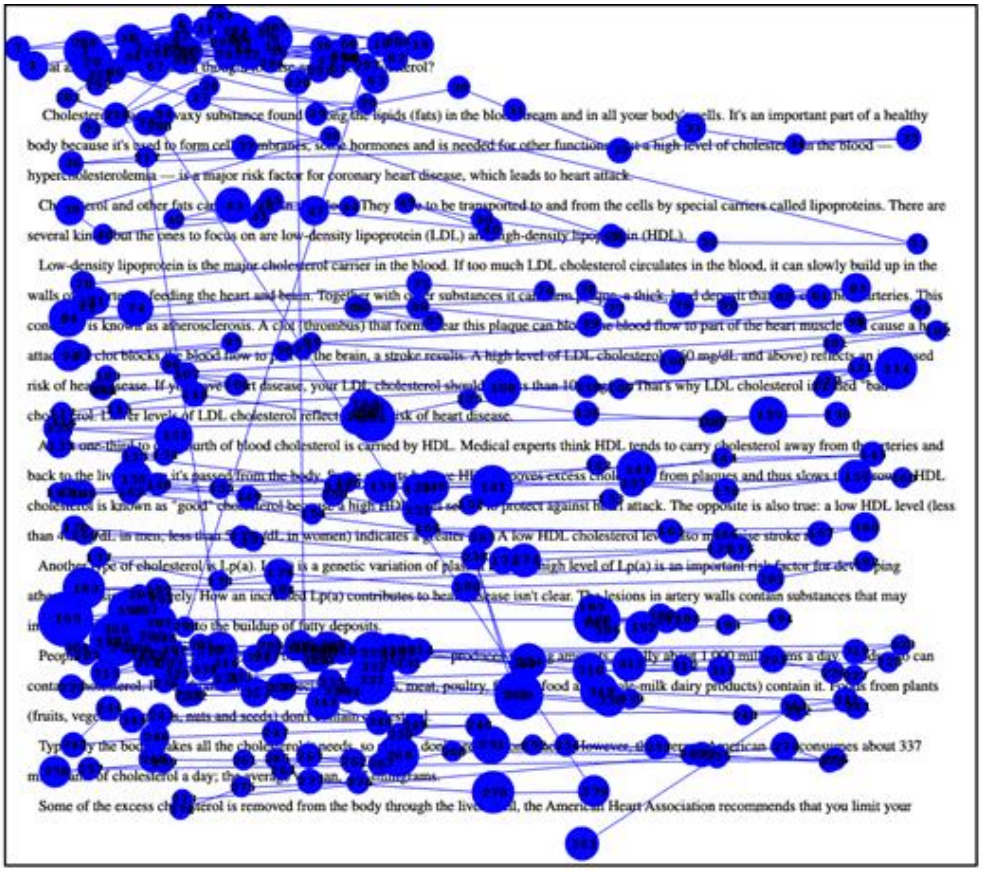


Figure 1 Gaze plot results of the question and the answer using Tobii Eye Tracking System

hot spot image showing where the eye gazed and how long the eye gazed at a particular spot, represented in intensity of color (see Figure 2). The system also allowed me to define “Areas Of Interest” (AOI) by defining certain rectangular-shaped areas in the document.

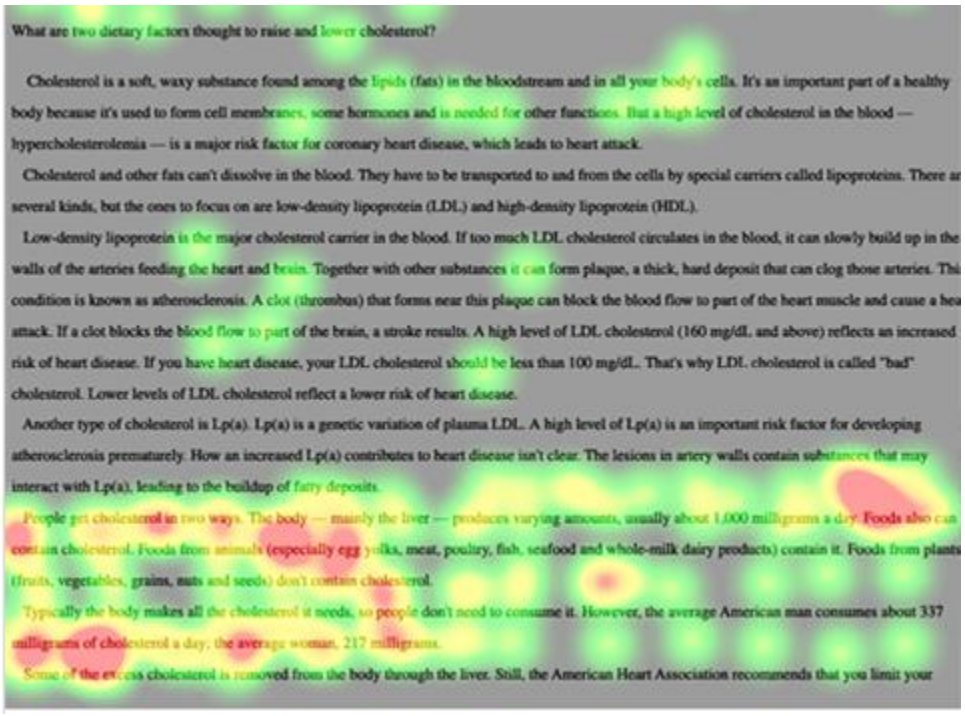


Figure 2 Hot spot image results of scanning for the answer to, "What are two dietary factors thought to raise and lower cholesterol?" using the Tobii Eye Tracking System

For the initial analysis I chose to put an AOI rectangle around each title, header, list (numbered or bulleted), and paragraph of the document. The system then produced a text file that contained the time and duration of each hot spot within an AOI for each subject (see Appendix B).

Because of the inconsistent answer results and the sheer amount of data for the longer documents, preliminary analysis was limited to the 10 two-page documents.

Analysis of gaze plots and hot spot images showed that there were three techniques subjects used to locate the answers. One technique subjects used was to move their gaze slowly throughout the entire document, indicating that they were most likely reading the document. With two-page documents, it was possible to read the

documents in a reasonable amount of time, although even in these documents people appeared to focus longer on (and thus were apparently reading) certain areas (as determined by the number of gaze hot spots in a particular area) and focused less on other areas (those areas having fewer gaze hot spots).

The second technique subjects used to peruse the document was to start at the top and move down, with hot spot gazes jumping randomly and quickly horizontally as they move vertically throughout the page (described as a "rough zig-zag movement" by McLaughlin when studying the techniques of scanners (1969)), and never really stopping their gaze for extended periods of time throughout the document. Surprisingly, while the gaze of these subjects didn't appear to stop long enough for them to gather much information, they often were able to correctly answer the question.

The third technique subjects used to scan through documents was a combination of the above two techniques: subjects would gaze lightly and quickly throughout areas in the document, then stop and have many gaze spots at other areas in the document. This is similar to what Taylor (1962) found with speed readers: his readers skipped some lines yet had multiple fixations on others. An example of this technique is visible in Figure 2. While the data from all three groups was used for analysis, the data from this group was clearly the most relevant to my task since their fixation points most clearly showed what areas subjects found most interesting while scanning for an answer to a question.

2.1.3 Analysis of Scanning Data

To determine exactly which AOIs subjects focused on most frequently, I used the rectangular AOIs defined around headers, titles, paragraphs, and lists, as explained

before. I used paragraphs for a couple of reasons. With the Tobii Eye Tracking System at the time of analysis only allowing me to define AOIs with a rectangular shape, I needed to pick logical AOIs. Sentences were not an option because they often started in the middle of one line and ended somewhere in another line, thus not falling nicely into a rectangular shape. Lines were also not an option because, even after calibration, the Tobii Eye Tracking System is only accurate within 2 mm, which is often about the space between two lines in my documents. Thus I'd have no accurate way of knowing exactly where the subjects intended to focus based on the eye tracking gaze point information. Finally, if the AOIs had been defined as a relatively small area, I still might not know exactly what was causing an individual to focus on a particular area. It has been shown that groups of information within 5 degrees of visual angle can be perceived in a single eye fixation (Tullis,1986). Thus, while the eye gaze may fall within a particular place, to more accurately gauge what cue resulted in subjects focusing on a particular area, I needed to look at a larger block of text. Thus I chose to use paragraphs, which allowed me to evaluate areas of text users focused on, thus ameliorating the problem of imprecise eye tracking information and peripheral gaze information. I counted the number of gaze points (or focus points) in each AOI across all subjects using the 10 two-page text file produced by the eye tracker for each subject for each question/document pair.

In looking at what information individuals focused on while scanning, I found that the focus areas of individuals could be classified into two categories: Areas focused on because of physical cues; and areas focused on for reasons other than visual cues and thus most likely because of their content and its relationship to the question.

2.1.3.1 Physical Cues

Because I was interested in what visual scanners focused on other than physical cues, I purposely chose articles for scanning that had very few physical cues in them. None had images in them. As mentioned before, a few had subtitles or lists in them, but many consisted of one title, possibly an author, and then text in paragraph form. However, in analyzing data, I found that, when present, individuals did focus on the title and subtitles that occurred in the documents (as exhibited in Figure 1, page 12). Subjects also frequently focused on the first paragraph[2] or paragraphs of a document. There was less of a tendency, but still a trend for focusing on the first paragraph on each page. Interestingly, although a few subjects focused on the first line of each paragraph, this was not a common practice. This is significant because it is a technique available to users of screenreaders, yet it clearly does not give users of screenreaders the same type of information that visual scanners get when scanning through a document.

Clearly subjects have learned to pick up on visual cues because they provide needed information. This correlates with Hovy et al.'s (1997) Optimal Position Policy (OPP), in which they found that certain sentence positions within a document were more likely to contain topic-relevant information. For instance according to them, in the Ziff Davis corpus, the order in which the most relevant sentences occur is: Title 1; Paragraph 2, Sentence 1; Paragraph 3, Sentence 1; Paragraph 4, Sentence 1; Paragraph 1, Sentence 1; Paragraph 2, Sentence 2; etc. It is likely that people learn approximately where physically in a document the most useful information is located. Thus the

---

[2] The first paragraph was the second-most focused on paragraph in four of the 10 two-page documents (second only to the paragraph with the answer to the question), and in all but one of the two-page documents, it was one of the top four most focused on paragraphs.

system I created includes options for giving information on where topologically within the document the user is, and thus should aid nonvisual readers in the use of these topological cues.

2.1.3.2 Semantic Cues

Many areas focused on did not have notable physical features that may have attracted attention. I specifically wanted to look at these areas. My conjecture was that these AOIs were focused on by subjects because of their semantic relationship to the question. Indeed, I did find evidence of this. Results showed that for seven of the 10 two-page documents and questions, subjects focused most on the paragraph containing the answer to the question[3]. As an example, one of the questions used in the study was,

> *"How do people catch the West Nile Virus?"*
>
> The paragraph with the most gaze points for the most subjects was:
>
> *"In the United States, wild birds, especially crows and jays, are the main reservoir of West Nile virus, but the virus is actually spread by certain species of mosquitoes. Transmission happens when a mosquito bites a bird infected with the West Nile virus and the virus enters the mosquito's bloodstream. It circulates for a few days before settling in the salivary glands. Then the infected mosquito bites an animal or a human and the virus enters the host's bloodstream, where it may cause serious illness. The virus then probably multiplies and moves on to the brain, crossing the blood-brain barrier. Once the virus crosses that barrier and infects the brain or its linings, the brain tissue becomes inflamed and symptoms arise."*

This paragraph contains the answer to the question, yet it has very few words in common with the question. The words it does have in common with the question,

---

[3] For one of the documents, the paragraph containing the answer was tied with another paragraph as having the highest number of gaze points.

‘West Nile Virus’, are the topic of the document and occur fairly frequently throughout the document, and thus cannot account for subjects' focusing on this particular paragraph.

The subjects appear to have made semantic connections between the question and the answer that cannot be explained by simple word matching or even synonyms, hypernyms and hyponyms. This suggests that when scanning, subjects were able to make the semantic connections necessary to locate question answers, even when the answer was of a very different lexical form than the question.

Other areas of text focused on also appear to have a semantic relationship with the question. For example, with the question,

“Why was Monet’s work criticized by the public?”

the second most frequently focused on paragraph was:

> *“In 1874, Manet, Degas, Cezanne, Renoir, Pissarro, Sisley and Monet put together an exhibition, which resulted in a large financial loss for Monet and his friends and marked a return to financial insecurity for Monet. It was only through the help of Manet that Monet was able to remain in Argenteuil. In an attempt to recoup some of his losses, Monet tried to sell some of his paintings at the Hotel Drouot. This, too, was a failure. Despite the financial uncertainty, Monet’s paintings never became morose or even all that sombre. Instead, Monet immersed himself in the task of perfecting a style which still had not been accepted by the world at large. Monet’s compositions from this time were extremely loosely structured, with color applied in strong, distinct strokes as if no reworking of the pigment had been attempted. This technique was calculated to suggest that the artist had indeed captured a spontaneous impression of nature.”*

Of the 30 subjects who scanned this document, 15 had their largest number of focus points in this paragraph, making it the second most focused on paragraph in the document, second only to the paragraph that contained the answer (most focused on by 21 of the subjects). The above paragraph occurred within the middle of the second

page of the document, with no noticeable physical attributes that would have attracted attention. Upon closer inspection of the paragraph, there are references to "financial loss," "financial insecurity," "losses," "failure," and "financial uncertainty." One also sees, "morose" and "somber" and even "had not been accepted by the world at large." Subjects appear to be making a connection between the question topic, Monet's work being criticized by the public, and the above terms. Intuitively, people do seem to make this connection. Yet the connection being made is not straightforward and cannot be replicated using the direct semantic connections that are available via WordNet (Felbaum, 1999), which provides basic relationships between words including synonyms, hypernyms, hyponyms, antonyms, etc. along with short definitions. Indeed, the relationships made are more similar to Hovy and Lin's (1997) Concept Signatures created by clustering words in articles with the same editor-defined classification from the Wall Street Journal. Clearly the system I created needed to be able to replicate these connections automatically.

Upon further examination, I found other paragraphs that were focused on by subjects for reasons other than their physical appearance or location, yet their semantic connection to the question is even more tenuous. For instance, when scanning for the answer to the question,

"How does marijuana affect the brain?"

the third most frequently focused on paragraph (third to the paragraph with the answer and the first paragraph) was,

> *"The main active chemical in marijuana is THC (delta-9-tetrahydrocannabinol). The protein receptors in the membranes of certain cells bind to THC. Once securely in place, THC kicks off a series of cellular reactions that ultimately lead to the high that users experience when they smoke marijuana."*

While this paragraph does appear to have loose semantic connections with the question, the connections are less obvious than paragraphs that follow it, yet it is this paragraph that subjects chose to focus on. The paragraph is the third to last paragraph on the first page, so its physical location cannot explain its attraction to subjects. If, however, one looks more closely at the previous paragraphs, one sees that the first paragraph deals with definitions and alternate names for marijuana (with no semantic links to the question), and the second and third paragraph deal with statistics on people who use marijuana (again, with no semantic connection to the question). The fourth paragraph, the one focused on, represents a dramatic semantic shift towards the topic of the question. Intuitively it makes sense that individuals scanning through the document would pay more attention to this paragraph because it seems to represent the start of the area that may contain information related to the answer, not to mention conveying topological information about the general content of the document.

2.1.4 Experiment Limitations

While the data collected in this study implies that semantic connections are made, the scope of this experiment must be kept in mind. The vast majority of data (and the data examined most closely) was the data associated with two-page documents and their corresponding question. This was done for two reasons: to keep the length of the study for subjects manageable, and to keep the amount of data for each question-document pair at a manageable level. However, it would be somewhat arbitrary to assume that visual scanners scan longer documents in the same manner in which they scan shorter documents when searching for answers to complex questions. It is entirely possible that there are other factors that are engaged when a person is scanning through a longer document such as a chapter in a book.

2.1.5 Experiment Summary

Data collected from these experiments suggest that within shorter documents subjects do make and scan for semantic connections. Subjects paused on information that directly answered the question even when the answer was spread out over a few sentences and did not contain words in the original question. They also focused on other content within the document that was semantically related to the question. While physical attributes of text do attract the attention of scanners, and thus the system must include methods for accessing this data as well, it is clear that in order to create a successful scanning device that conveys information similar to what visual scanners get when scanning for the answer to a question, a method needed to be developed for automatically generating loose semantic connections and then using those semantic connections to locate the text the visual scanners considered to be relevant within the document.

Chapter 3

NLP TECHNIQUES TO IDENTIFY RELEVANT TEXT

In order to automatically generate the semantic connections that might explain what caught the eye of the visual scanners, causing them to gaze longer and more frequently in some places, I explored ways of recreating the connections between the words in the question and the words in areas of text. For this I first investigated existing Natural Language Processing (NLP) techniques.

3.1 Question Answer Systems

My goal was to recreate the semantic connections people made when scanning for the answer to a question. While this is a different goal than most Question Answering Systems, whose goal is to search a document for the answer to a question, the eye tracking experiments did show that visual scanners often gazed more frequently in the paragraph containing the information that answers the question. In addition, they were able to actually correctly answer the question after scanning. Thus I first looked at the techniques used by existing Question Answering Systems to examine the types of semantic connections they make.

3.1.1 Related Research

Question Answering Systems take as input questions in their natural language form and search through a document or a (potentially very large) set of documents to locate the answer. Answers may then be retrieved from the returned documents. Depending on the system, the Question Answering System can either return text

excerpts containing the answer, a summarized answer, or a link to the answer within a document. There are two categories of Question Answering Systems: Restricted Domain and Open Domain.

3.1.1.1 Restricted Domain Question Answering Systems

With restricted domain question answering systems the domain is contained within one or a few known topics. With only one or a few domains, specific ontological information about the domains can be built into a model of the domain so that more complex reasoning can be implemented when attempting to answer a question. Early restricted domain systems developed a hand coded database of information on a particular topic and focused on transforming the user's question in natural language form into queries that could be used to query the database for the answer (e.g., BASEBALL (Green et al., 1961), which answered questions about a particular year of baseball, and LUNAR (Woods, 1997), which answered questions on the Apollo moon mission's rock analysis). Systems today focus on the formation of a knowledge representation including the ontological concepts and relations between words. Questions, often complex in nature, are transformed and answered using the specific knowledge representations and reasoning mechanisms built into the system. Examples of these types of systems include Zajac (2001) and Barker et al. (2004).

Restricted Domain Systems do connect complex questions (posed in natural language) to the answer in a text document or document. However the connections are made within a limited data set. Many of these systems develop an external knowledge base, specific to the domain, that is utilized in answering questions. However, because the system I developed needed to work over an unrestricted domain (and a constantly changing set of domains), building the ontological relations over all domains that is

necessary for the deeper reasoning was impossible. Word sense disambiguation, which is usually easier within a specific domain, is difficult when reasoning over unrestricted domains. Specific terminology and word relations are impossible to code ahead of time for all possible relations. Thus the Restricted Domain System solutions were not adequate for my goals.

3.1.1.2 Open Domain Question Answering Systems

With unrestricted (Open) Domain Question Answering Systems, the domain is not known ahead of time, nor is it limited to one or a few topics. Specific ontological relationships cannot be built into the system, nor can domain-specific vocabulary and word senses. A good deal more work is being done with Open Domain Question Answering Systems today because of their widespread use in search engines as well as conferences such as the Text REtrieval Conference (TREC), an annual information retrieval competition started in 1999 (Vorhees) with a question answering track. Because most systems are only capable of surface reasoning over open domains, the focus of these Question Answering Systems has largely been on answering simple, fact-based questions, e.g., "How tall is Mt. Everest?", or "Where was Abraham Lincoln born?" Because of the TREC competition, later Open Domain Question Answering Systems have focused on other types of questions included in the competition, including list type questions (e.g., "Name famous people who have been Rhodes scholars") and definitional 'other' questions (e.g., "What is the Islamic counterpart to the Red Cross", "What is 'autism'?").

The research focus of these Question Answering systems can be divided into two areas: Converting the question into a query to be used for retrieving potential

answers and/or documents with the answer from a set of documents; and retrieving and rating the potential answers.

3.1.1.2.1 Question Transformation

In transforming the question into a query, techniques range from simply using the question's non-stop words as query terms, to reformulating the question as a potential answer string, to classifying the question based on the expected answer type. Most Open Domain Question Answering Systems use a combination of these techniques to form their queries.

Kwok et al.'s (2001) MULDER system created a progressive series of queries, which were submitted simultaneously to the World Wide Web. At its simplest, the query was a set of words, measured in importance by IDF (Inverse Document Frequency), from the question. The most specific query was a partially quoted sentence that was a reformulation of the question (e.g., the question "Who was the first American in space?" would be reformulated into the string, "The first American in space was"). Queries in between in specificity include using queries with noun phrases in quotes (which meant that the phrase must be matched exactly and in its entirety) and expanding the vocabulary to include attribute nouns of adjectives (e.g., 'tall' – 'height') using WordNet. Like other systems, another key component of the MULDER system was its question classification. MULDER classified questions based on expected answer type (e.g., Nominal, Temporal, Numerical). For instance, "How many…?" would be classified as Numerical, whereas "Where…?" would be classified as Nominal and "When….?" would be classified as Temporal. The object of the verb was used to determine the type in questions that couldn't be determined using just the first word(s). Using WordNet, hypernyms of the object were traversed until "measure"

or “time” were found, and the question was classified appropriately. This classification was used to identify potential correct answers.

Brill et al.’s AskMSR System (2002) also formed multiple queries to send to a search engine, although their system chose to take advantage of the vast amount of information on the Web, leading to redundancy in the appearance of a question’s answer. In reformulating the query, AskMSR classified questions for one of 7 possible answer types, each with its own rewrite rule. Rewrite rules were simple string manipulations, with no parsing involved. The query rewrite rules resulted in tuples that contained the string to be matched, whether the answer was expected to the left or right (or no preference) of the string, and the weight of the likelihood of the answer occurring based on this tuple. So, for instance, if the question was “Who killed Abraham Lincoln?”, the query rewrite rules would result in the string, “Abraham Lincoln was killed by”, with the expected answer location being right, and a weight higher than the weight of the lower precision query, “Abraham”, “Lincoln”, “killed”.

Srihari et al. (2000) focused on creating templates with a set of keywords from the question and what they referred to as “asking points,” or Named Entity types that describe the expected answer type. The asking points included a wide range of types, including person, organization, location, time date, money, percent, duration, frequency, age, number, fraction, decimal, ordinal, math equation, weight, length, temperature, angle, area, capacity, speed, rate, product, software, address, email, phone, fax, telex, www, name, etc. The system also included subtypes and predefined notions like REASON (e.g., for the question “Why did David Koresh ask the FBI for a word processor?”, the asking point does not match any other Named Entity type definitions, and thus becomes the notion REASON).

Hovy et al.'s (2000) Webclopedia system focused on parsing the question into noun phrases, nouns, verb phrases, verbs, adjective phrases, and adjectives, and then assigning each of these query terms a weight based on how often their type occurred in the corpus used, on the length of the query term, and on the frequency of occurrence of the query term in the document. Query terms were expanded using WordNet's synsets, and the returned terms were placed in a Boolean expression (e.g., high & school | senior & high & school | senior & high | high | highschool).

3.1.1.2.2 Answer Retrieval

In Question Answer systems, once a query has been formulated, documents potentially containing the answer must be retrieved and, in most cases, the text segments containing the answer must be located. In most systems, the formulated query or set of queries is submitted to a corpus of (potentially indexed) documents, often using a search engine, and documents that have the highest matching score to the query are returned as documents that potentially contain the question answer. While some systems search through a small corpus for documents related to the query (e.g., the TREC corpus) for the answer (Hovy et al., 2000; Katz et al., 2005; Galea, 2003; Srihari et al., 2000), others search the World Wide Web for documents with the answer, formulate an answer using the WWW document information, then search through a smaller corpus for a matching answer (Kwok et al., 2001; Brill et al., 2002; Buchholz, 2001).

Open Domain Question Answering Systems that use the World Wide Web as a knowledge source have a vast base of potential answer documents. Thus systems can rely on the fact that most likely numerous documents will contain the answer to a question. With many documents containing the answer, the chances of finding a

matching answer to a posed query are high. Thus these systems can afford to miss ill-formed answers or answers that are phrased in a way that does not match the query. In addition, because of redundancy of the correct answer, if an incorrect answer appears as a match to the query, it will most likely be filtered out in favor of the more frequently occurring correct answer.

Once a refined set of potential answer documents has been defined, the systems must then rank and retrieve passages that contain the answer. Text segments are ranked by how well the segment type matches the query type, by the number of important keywords the text segment contains, and by how close they are to one another in the segment. In Hovy et al.'s Webclopedia (2000), the query terms contained question words and phrases and their synonyms. The system first returned a set of documents ranked by giving each matching term 2 points and each matching synonym 1 point. Those documents were then segmented using Hearst's TextTiler (1993), the segments were ranked using the above scoring method, and the top 100 segments were parsed for matching question and answer type.

Abney et al. (2000) gave scores to text segments, which it considered to be a sentence and its two adjacent sentences. Each sentence received a matching score based on the number of words matching query words from the question weighted by IDF. A passage received a matching score based on the formula 1/4 Si-1 + 1/2Si + 1/4Si+1. Each passage was then ranked for its likelihood of containing the answer based on whether the query's type and the answer's type matched, and the frequency and position of the entity matched to the type in the answer.

Srihari et al.'s (2000) system evaluated a set of documents retrieved from the query for information extraction. The returned documents were tokenized and tagged

for part of speech and type (e.g., person, location, time date, etc.). Then text was ranked first based on how many keywords are in a sentence, then on the order of the keyword occurrence in the sentence compared to the order in which it occurs in the question, and then on whether the key verb or a variant of the key verb is exactly matched. The type is also matched.

Brill et al.'s AskMSR (2001) focused on taking advantage of the redundancy available via the web to locate answers. Summaries returned as a result of a query are each made into 1-gram, 2-gram, and 3-grams, each weighted by the weight of the query that retrieved it and the sum of those weights across all summaries containing the n-gram. The more summaries that contain the n-gram, the higher the weight given to that n-gram. Then an answer tiling method is applied such that n-grams are merged if part of the n-grams overlap. The weight of the new n-gram is the maximum weight of the n-grams making up the new (longer) n-gram. The highest weighted n-grams are always checked to see if they can be merged with the new n-gram first. In this way, the answer that occurred most frequently gets merged into the potential answer tiles, and thus the redundancy of answer occurrence in summaries returned from searches is taken advantage of.

Kwok's MULDER (2001) further expanded on taking advantage of redundancy by clustering similar answers into one group with a higher rank. Thus, for the question, "Who was the first American in space?", "Shepherd", "Alan B. Shepherd", and "Alan Shepherd" would be clustered together, and the cluster group would rank higher than "John Glen", which in some web documents could in all ways also match the query, but would be an example of an incorrect answer located on the Web and thus should occur much less frequently.

3.1.1.2.3 External Knowledge Sources

Later systems used external knowledge sources to build in general ontological information for use in question answering. Agirre et al., (2000) explored using the World Wide Web to form topic signatures, which consisted of topically related words. The goal was to create a collection of related words for each concept in WordNet, which could then be used for word sense disambiguation, among other things. WordNet was used to build queries for each word sense in WordNet. The queries were then used to retrieve a collection of documents from the World Wide Web. To eliminate documents that might relate to other senses of the same word, queries were constructed such that they did not contain query words that were from another sense of the word. The system used all documents retrieved. Documents returned were then analyzed for words and frequencies. Word frequencies were compared for each set of documents resulting from the different queries, and those with a high distinctive frequency for a particular document collection became part of the topic signature for that WordNet concept.

Katz et al., (2005) used Wikipedia as an external knowledge base for list questions. In Wikipedia, certain articles can be considered full-article lists because the entire article is a list. The Katz system took the query terms from the question and attempted first to match them to list names from Wikipedia article titles. Subtitles and redirection links under the title were considered possible synonyms (the term 'synonym' was used loosely) for query matching purposes. If matches were found between the question and the titles, etc., the potential list items found in the Wikipedia article were treated as possible answers to the question.

Because of the growing focus on answering questions that are less straightforward and may have implicit rather than explicit answers, Clark et al. have

an ongoing process that involves extending the general ontological knowledge available via WordNet (2008). Their extensions involve including morphosemantic links (e.g., the verb 'build' is linked to the noun 'builder', which is the agent of the verb), purpose links (e.g., 'gun' exists in both noun and verb form – a 'gun'(n.) is for 'gunning'(v.)), general world knowledge that can be built from a word's glosses and sentences as well as other word's glosses and sentences that contain the word, and general, or "core" world knowledge about things like space, time, events, communication, etc. that must be hand encoded.

Banko et al., (2007) developed a system that allowed for more complex relational queries. His system, called TextRunner, indexed a set of relational tuples over the World Wide Web for query answering. The system first parsed training data into noun phrases, then traversed the parse tree to extract relations between the noun phrases for each question. The resulting tuples (entity-relation-entity, where the entities are the noun phrases) were converted into feature vectors, which go into a Naïve Bayes classifier. Once the Classifier had been trained, new text was first tagged for part-of speech, then chunked into noun phrases, with the relations considered to be the text between the noun phrases. The relations were tagged for probability, then indexed and stored. When indexing the web, the tuples were stored over a pool of machines so that queries could find relevant relations quickly.

Soricut et al., (2006) developed a preliminary Open Domain Question Answering System that handled questions that were also not restricted to simple, factoid questions. They used as an external knowledge source a set of approximately 1 million question-answer pairs retrieved from FAQ pages on the web. The system used a statistical chunker trained on the answers of FAQ knowledge base (to counter for

any stylistic gaps between the questions and answers). The chunker chunked the natural language question into 2 and 3 word chunks that were used as query terms. The query terms were used with a search engine on the Web, and the first N documents returned were segmented into sentences and evaluated for potential answers. It was assumed that answers could be located within 3 consecutive sentences (because 3 sentences were the average number of sentences used in answers in the FAQ knowledge base). A statistical translation model was used to relate questions to potential answers, in which the probability of each answer being the answer to a question is calculated using probability models computed using the FAQ knowledge base.

3.1.2 Issues with Current Open Domain Question Answering Systems

Open Domain Question Answering Systems still largely deal with simple, factoid questions or questions that can be answered in a limited number of sentences. A great deal of effort is being placed on parsing and reformulating the question into a potential answer form. As has been found by Soricut et al. (2006), reformulating complex questions more often hurts performance than improves it. Thus existing systems are not well equipped to deal with complex questions, in which the relationship between the question and the answer may not be straightforward in nature. Many assume that if an answer is ill-formed or atypically formed, the redundancy of a vast database will eliminate the need for detecting a relationship between a question and related, but not necessarily well-formed, text within a document. Existing systems do not take into account the possibility that information pertinent to a complete understanding of the question very well may occur in noncontiguous parts throughout a document. If, however, the system is limited to locating an answer within a few or

even one document, then it cannot afford to miss atypically formed answers, answers that are longer than a few sentences, or answers that occur over noncontiguous text. Equally, these systems do not take into account the possibility that a user might need knowledge not directly answering the question, but nevertheless related to the question, in order for the user to develop a more in-depth understanding of the answer to a question. To make these less direct connections, a system must rely on external ontological knowledge across domains. Some systems today are building and using external knowledge bases, but creating a knowledge base containing ontological and general relational knowledge across all domains is challenging. According to Brill et al. (2000) "Given a source that contains only a relatively small number of formulations of answers to a query, I may be faced with the difficult task of mapping questions to answers by way of uncovering complex lexical, syntactic, or semantic relationships between question string and answer string. The need for anaphor resolution and synonymy, the presence of alternate syntactic formulations, and indirect answers all make answer finding a potentially challenging task."

3.2 Text Summarization

Another area of research from which I can seek relevant inspiration is text summarization, especially single-document summarization and query-biased summarization. Text summarization is the act of taking a document or set of documents and creating either an extracted or abstracted summary significantly shorter in length than the original document, either based on a topic or query, or purely based on the content of the document(s). Usually with multi-document summarization, a topic or query is used to retrieve a set of documents, and the documents are often

ranked in terms of relevance, then a summary is created, often using the topic or query as a reference point. If no topic is provided, the topic must first be identified.

3.2.1 Related Research

In their topic-driven text summarization system, Carbonell and Goldstein (1998) took both relevance and redundancy into account when determining the importance, or rank, of a document. Each document returned from a query sent to a search engine was selected to be included in the summarization by measuring its weighted similarity to the query minus the weighted maximum similarity of the document to each of the set of documents already selected for summarization. They called their measure Maximal Marginal Relevance (MMR) measure. The documents with the highest MMR measures were selected to be part of the set of relevant documents, with the process repeating until a set number of documents were selected. Originally, the weights get set so that a document's similarity to the query start higher than the weight of its similarity to other selected documents, and as the process continues, the weights progressively reverse. Once documents were selected, they were segmented and MMR were applied to the segments. The top segments were selected and used to create a summary.

Hovy et al.'s (1997) SUMMARIST system used position to create summaries. They came up with the Optimal Position Policy (OPP)[4], based on the finding that certain sentence positions in a document were most likely to contain the most topically-relevant information. So for different genres of documents they generated an OPP, or optimal sentence order for garnering the most relevant information. Hovy et

---

[4] Previously discussed in Chapter 2

al. were also interested in identifying the topic of an article by fusing a set of semantically related keywords into one unifying concept. They used WordNet to find relations among keywords. They also developed Concept Signatures, in which they related sets of words into one concept using as an external knowledge source the Wall Street Journal. The Wall Street Journal has 30,000 articles, each with a classification concept identified by editors. Hovy et al. counted occurrences of each content word in a set of articles with a specific classification, weighted them using TF/IDF, selected the top 300 terms, and used the classification as the head signature topic for the terms.

Later Hovy et al. (2005) and Zhou et al. (2006) introduced the concept of Basic Elements for creating summarized answers to more complex questions. They defined a Basic Element (BE) as a head-modifier-relation triple, with the head being a noun, verb, adjective, or adverb phrase that acts as a major syntactic constituent and its relationship between some dependent entity. So "two Libyans were indicted…" would become: Libyans|two|nn and indicted|Libyans|obj. Each sentence in a multi-document set was broken into BEs and each BE was ranked by giving it a likelihood ratio indicating its importance in the document set to be summarized. Each sentence was then ranked according to the score of its BEs. The top sentences were added to the summary by first adding the top ranked sentence. Then each subsequent sentence was checked for overlap ratio R based on its similarity of the BEs with those already selected for the summary to prevent redundancy. Positional information was also considered in the weighting of sentences to be included in the summary. In this system the query was used as the topic. The query was tagged with part-of-speech tags and expanded using synonyms for the nouns and verbs using WordNet. Terms were then

weighted using inverse term frequency using the Wall Street Journal, and the top weighted terms were used as the topics.

Otterbacher et al. (2006) developed a system for retrieving sentences in response to a factual-based question from a set of complex news articles about multi-event stories that occurred over time. Their system created a graph, with each sentence being a node in the graph and edges occurring between nodes (sentences) when the cosine similarity between the two sentences exceeded a threshold. Nodes were given a degree based on how much information the sentence node had in common with other sentences. So sentences with a lot in common with other sentences, and thus most likely to be related to the topic of the documents, would become centrally located in a graph with a higher degree. Then nodes were ranked based on their eigenvector centrality, or how well connected the node was. Once the graph had been created, the job of locating sentences relevant to the question began. Sentences were retrieved using both their measure of similarity to the query (relevance) and their measure of similarity to other sentences deemed relevant to the query. In the documents, all sentences were stemmed and word values were determined using IDF. The question was also stemmed and stopwords were removed. Then the similarity of a sentence and the question was determined by comparing the words in the sentence to the words in the question weighted by the word's IDF. Once they found a sentence that appeared very relevant to the query, they included sentences that appeared to have less relevance to the query but were highly connected to the sentences that were highly connected to the query. The assumption was that even though a sentence may not appear to be connected to the query, if it was connected to another sentence that was highly connected to the query, then it very well might contain pertinent information as

well. Thus each sentence was determined to be related to the query both by its relation to the question and its relation to sentences already chosen to be in the cluster of answer sentences. Initially, the relevance of a sentence to a question was weighted much more highly than a sentence's relationship to other sentences in the answer sentence cluster, and as sentences were chosen, that weighting shifted to weight a sentence's relationship to the answer sentence cluster more highly.

3.2.2 Issues with Text Summarization Systems

Text summarization systems lend interesting ideas about locating and connecting related text segments within documents. The SUMMARIST system went so far as to identify conceptually similar words using concepts identified by the Wall Street Journal. My system needed to locate text within a single document that may or may not contain a clearly identified topic that relates directly to an area or areas of text within a document. Unlike most text summarization systems, however, I needed to identify areas of text within the document that may not contain any words directly related to the query. While visual readers are able to scan through a document and focus on areas of text within the document that have a loose semantic connection to the question, including the answer to the question, current text summarization systems require direct word overlap at some level, either between the question or topic and the text, or between areas of text within the document, in order to identify relevant text. In contrast, my experiments showed that visual scanners focused in portions of text with no direct word connection with the question and my system needed to identify these.

My system is unique in that it has as its goal not to answer a question or create a summary, but to return information visual scanners glean while scanning through a document when answering a question. Questions posed to the system will range from

simple to complex in nature, and the related text (including the answer) must be found within a single document, regardless of the form the answer takes. Questions can be on any topic. With complex questions, it is rarely possible to categorize the type of question (and thus the expected answer type). Intuitively, it appears equally useless to attempt reformulation of the query for pattern matching. This intuition is born out by Soricut and Brill (2006) who stated that in their study reformulating complex questions more often hurt performance than improved it. Note that while the goal of my system is to give the user all information related to the question as well as a general topology of the document, especially as related to the question, similar to the information a visual scanner gets when scanning, the system should also give the user the control to use the information to decide where to read more thoroughly and where to skip over . My system must also work in real time. Thus, while I was able to utilize some methodology developed for text summarization, I needed to expand on those techniques in order to give the users of my system the ability to generate their own topology of a document, focus on all or most of the text in a document relevant to a complex question regardless of the form the text takes, and allow them to stop and focus or continue traversing the document's relevant text, all in real time.

Chapter 4

SEMANTIC CONNECTIONS

In order to identify text relevant to a complex question within a single document, I needed to find a way of matching the question's words to specific areas of text in the document. I chose to use as a unit of text the paragraph for a number of reasons. While I had the choice of matching words, sentences, paragraphs, and possibly entire sections and chapters, the most logical unit size was sentences and paragraphs. Simply matching words wouldn't give enough information about surrounding text, and thus wouldn't indicate how well the text area corresponded to the topic of the complex question. Sentences were a more logical unit (and were frequently used in many of the question-answering systems and text summarization systems), but I was especially interested in areas within the document in which a number of sentences that were semantically related to the question were located. Logically, an area of text with a number of sentences highly correlated to the question is a paragraph. In addition, while paragraphs should be consistent in their topic (and thus consistently more or less related to the question), larger areas of text like sections or chapters may not be as consistent in terms of the topic(s) they're covering. Finally, the information gained from the eye tracker study (discussed in Chapter 2) showed the paragraphs most focused on by visual readers. Thus I had data on not only what paragraphs held the answer (or at least text most related to the answer), but I also had data on which paragraphs visual readers focused on most (suggesting a relation between those paragraphs and the question in the minds of the visual scanners). Thus I chose to focus on paragraphs, and especially on the relationship of the sentences that make up the paragraph to the complex question.

Paragraphs also serve as a useful unit for conveying topological information to a user (e.g., the 5th paragraph of a document), thus making it useful in the creation of the final system. Indeed, many screenreaders allow users to traverse a document by jumping both backwards and forwards through the first sentence of paragraphs, making paragraphs a topological unit users of assistive reading technology are familiar with.

For my system I wanted to take a complex question and a related text document and identify and rank the paragraphs whose content was most relevant to the question. In order to make the connection between the complex question and the paragraphs, I borrowed and expanded on techniques used in Question Answering Systems and Text Summarization Systems and used the results to rank the paragraphs according to their relevance to the question.

4.1 Baseline Connections

In order to determine how well my system needed to work, I wanted to establish how well baseline methodologies worked in identifying the question-relevant paragraphs within a document as determined by the eye gaze experiments. I looked at Direct Question Word Matching, and then expanded on that by incorporating Synonyms, Hypernyms, and Hyponyms (SHH).

4.1.1 Direct Question Word Matching (Baseline)

For the baseline I took the question, and directly matched the words in the question to the text in the document as described below. This was one method used in many open-domain question-answering systems for simple questions (Hovy et al.,

(2000), Abney et al. ,(2000), Brill et al., (2001), Kwok (2001)). I matched the question's words to the document paragraphs using the method described below.

4.1.1.1 Direct Matching Methodology

For the baseline I created a set of search terms using all the nonfunction words in the question (function words include words such as, "the", "and", "a", etc.). I counted the occurrence of each search term in each paragraph in the document associated with the question. Each of the search terms was weighted using a variant of TF-IDF (Term Frequency/Inverse Document Frequency) (Salton and Buckley, 1988). TF-IDF is a method used for calculating the informational importance of a word to the meaning of a document in a corpus. For example, if a document contained the words "any" and "hematocrit", you would most likely want the word "hematocrit" to be considered more seriously in trying to determine the meaning a document than you would the word "any" because "any" occurs much more often (i.e., in many documents and thus in many semantic contexts) and thus is much less indicative of the semantic context. As proposed by Salton and Buckley, TF-IDF calculates the relative importance of a word as follows: The TF term is calculated by counting the number of times a word occurs in a document normalized (divided) by the total number of words in the document. The IDF term is calculated by taking the total number of documents in the corpus, and dividing it by the number of documents the word occurs in. This is normalized by taking its logarithm. The TF and IDF are multiplied to give a word's relative weight, or importance in ascertaining the topic(s) of a document in a corpus.

For my purposes I wanted to adjust the weights of the search term words so that those dealing with the topic of the document had less weight than those dealing

with the topic of a particular paragraph or paragraphs within the document. While IDF normally refers to the number of documents in a set of documents a word occurs in, with the assumption that the more documents a word occurs in, the less relevance the word has to the topic of a particular document, in my case I am only looking at one document at a time. If a word occurs in all or most of the paragraphs in the document, then it most likely is related to the topic of the document, but doesn't help in identifying paragraphs within that document that pertain more directly to the question. So, for instance, in an article on Monet with its associated question being, "Why was Monet's work criticized by the public?", I would want the search term, "Monet", which occurs in most paragraphs in the document, to hold significantly less weight than, say, "public", which occurs very infrequently in the document, in determining the relevance of a paragraph in terms of the question.

In order to account for a word's significance in identifying the local topic (i.e., the paragraph's topic) within a document versus the overall document's topic, I adjusted the TF-IDF weighting scheme (which I will refer to as "Document TF-IDF") as follows: In this system I am trying to determine the importance of various paragraphs in a document. Thus, the search is limited to one document on a particular topic or set of topics. Therefore the document itself becomes the equivalent of a corpus, and each paragraph in my document can be considered equivalent to a separate document in the corpus. So to determine the Document Term Frequency (D-TF), I used the count of the occurrence of each search term in a particular paragraph (W_n), and to determine the Document Inverse Document Frequency (D-IDF), I used the count of the number of paragraphs that search term occurred in.

To calculate a paragraph's score in terms of how well it relates to the question, the following formula is used. Given a document containing P paragraphs: $P_1 \ldots P_p$ and a set of n search terms, $W_1 \ldots W_n$ I score each paragraph as follows:

$$\text{Score}(P_j) = \sum_{i=1}^{n} \left(\frac{\text{CW}_i}{CWP_j}\right) * \log\left(\frac{P}{CPW_i}\right)$$

Where

CWi = number of times Wi occurs in Pj,
CWPj = number of words in paragraph Pj
P = number of paragraphs in the document, and
CPWi = number of paragraphs containing Wi

4.1.1.2 Direct Matching Results

As expected, the results of locating relevant paragraphs using this baseline method were poor. In none of the 14 questions and documents did this method identify paragraphs shown to be relevant to the questions through the scanning experiment, nor did it ever accurately identify the paragraph with the actual answer (see Table 1, below).

Because of the complexity of the questions and the answers, this was expected. It clearly shows that simple word matching of the question and answer is not sufficient in identifying relevant areas of text.

Table 1 Rankings of the paragraphs using the baseline method

Question/ Document	Baseline ranking of paragraph ranked highest by visual scanners	Baseline ranking of paragraph holding the answer	Total # paragraphs in Document
QD1-La	8	8	14
QD2-St	3	3	15
QD3-Wn	5/6**	6	12
QD4-Co	6	2	14
QD5-Ci	5	4	11
QD6-Ea	Unranked*	Unranked*	13
QD7-Mo	Unranked*	Unranked*	10
QD8-Pi	6	6	15
QD9-Ma	Unranked*	Unranked*	12
QD10-Me	6	4	25

**If a paragraph contained none of the search terms or only search terms that occurred in every paragraph, it got a ranking score of 0, or in essence was unranked.*
***For this question/document pair, the paragraph containing the answer virtually tied with a second paragraph as being the most gazed upon by visual scanners*

4.1.2 Synonyms, Hypernyms, and Hyponyms (SHH) Search Terms (Baseline)

Clearly my system needed to make connections between the question and relevant paragraphs in a way that is more intelligent than simple question word matching. Because using synonyms, hypernyms, hyponyms, and even antonyms is a common technique used in both Open Domain Question Answering Systems (Prager et al., 2001; Hovy et al. 2002, Kwok et al., 2001) and in Query Biased Text Summarization (Varadarajan et al., 2006; Chali et al., 2002) and is what is commonly meant by semantic connections, I wanted to look at how well using the synonyms, hypernyms, and hyponyms of the search terms would work in identifying relevant paragraphs.

4.1.2.1 SHH Methodology

For the SHH Methodology, I expanded on the set of search terms by including all synonyms, hypernyms, and hyponyms of all the baseline search terms (i.e., the nonstop words in the question). WordNet (Felbaum, 1998) was used for the expansion. I was able to include all synonyms, etc., without worrying about homonyms (or words with different meanings but the identical spelling). It is logical to assume that all search terms generated by the homonyms would have equally non-relevant meaning and should logically match equally throughout the document, thus becoming a nonissue. Each of the words in the expanded set of search terms was weighted with the Document-TF-IDF weight, as described previously. Each paragraph was again given a matching score, and ranked according to its score.

4.1.2.2 SHH Results

Table 2 shows the results of using the SHH search terms.

Table 2 Rankings of the paragraphs using the SHH search term

Question/ Document	SHH ranking of paragraph ranked highest by visual scanners	SHH ranking of paragraph holding the answer	Total # paragraphs in Document
QD1-La	8	8	14
QD2-St	5	5	15
QD3-Wn	9/10**	10	12
QD4-Co	6	2	14
QD5-Ci	4	10	11
QD6-Ea	5	5	13
QD7-Mo	7	7	10
QD8-Pi	2	2	15
QD9-Ma	Unranked*	Unranked*	12
QD10-Me	4	7	25

**If a paragraph contained none of the search terms or only search terms that occurred in every paragraph, it got a ranking score of 0, or in essence was unranked.*
***For this question/document pair, the paragraph containing the answer virtually tied with a second paragraph as being the most gazed upon by visual scanners*

Interestingly, this method for identifying paragraphs relevant to the question was only marginally better than the baseline direct word matching method, and clearly not sufficient. In none of the question-document pairs did this method accurately identify the paragraph the visual scanners spent the most time on, nor did it make the connection between the question and the paragraph with the answer. It was clear that simply using synonyms, hypernyms, and hyponyms to make connections between the question and relevant text wasn't sufficient. To make these connections, I needed to develop a method that creates other, looser semantic links. Because the system needed to work with all questions and all documents, these semantic connections could not be defined a priori, and because this system has as a goal speeding up the process of accumulating information, these semantic connections needed to be made at the time of use of the system and with very little delay.

4.2 Making Loose Semantic Connections

In order to make these looser semantic connections, I needed to find a way of making the connection between words that are often discussed together. For instance, many would intuitively associate the words “dog” and “leash”. These words are clearly not synonyms, hypernyms, or hyponyms, yet are often discussed together. To make these connections I used the World Wide Web to find words frequently discussed at the same time that words in a question are discussed, and created a cluster of these related words. This cluster of words can be used to make the semantic connections between the question and text within the associated document.

4.2.1 Semantic Connections using the World Wide Web

The use of the World Wide Web to form semantic relationships isn’t new. To find the semantic similarity between two words, Matsuo et al. (2006) looked at the number of hits of each of the words as a single keyword search versus the number of hits using both words as the keyword search terms. The closer the two counts are, the more similar the words are. Chen et al. (2006) counted in each snippet of text returned from a Web search using word P the number of occurrence of word Q and vice versa. These values were used to compute the semantic similarity of words P and Q. The more snippets containing Q versus those that did not contain Q, the more semantically similar the two words were assumed to be. Bollegala et al. (2007) determined semantic relationships by extracting lexico-syntactic patterns from the snippets returned from a search on two keywords (e.g.,“‘x’ is a ‘y’”) and extracting the relationship of the two words based on the pattern.

Using the Web as a corpus of words has a number of advantages, including its sheer size and volume, the virtually complete coverage of all topics, and the fact that it

is constantly being updated with new terminology and semantic connections. For instance, not too long ago the word “Siri” (Apple’s “intelligent personal assistant”) would have produced few hits in a Web search, and those hits it did produce would have been obscure. Today in 2018, a Google search on “Siri” produces about 105 million URL snippets. Common semantic meanings are also constantly being updated to be current. The word “tweet” has a vastly different semantic meaning than it did 10 years ago, and a search of the World Wide Web results in an ordered list of URLs that reflect that shift.

The approach I developed is somewhat similar to Sahami et al. (2006) who used the snippets from a query word search to form a set of weighted words (weighted using TF/IDF) and then determined the semantic similarity of two keywords by the intersection of two word sets returned in those snippets. My work differs, however, in that rather than taking the snippet itself I take words from the text surrounding the snippet in the original Web page. Snippets are phrases found in the Web page which are most related to a Google search. They are intended to give the searcher some context to see whether the page is likely to be relevant. They typically contain the words that caused the Web page to be listed as a result of the search. The intuition is the snippet contains the words from the original question. For my system, I wanted to identify words that are typically discussed in conjunction with the question words. So my approach is to locate the snippet in the Web page’s text and take words surrounding the snippet as the larger context containing words that are potentially associated with the question.

4.2.2 My Approach to Identifying Relevant Words

In my approach I used the baseline search terms (i.e., the nonstop words from the question) as query terms for a search of the Web using Google (www.google.com). The search returned a ranked list of URLs and accompanying snippets of text.

For the top URLs returned, I took the snippet of text associated with it and divided the snippet up into a set of snippet phrases, using the "…" as the separator. For example, a search on " 'how' 'people' 'catch' 'west' 'nile' 'virus' " (from the question ("How do people catch the West Nile Virus?") returned the following snippet:

> *"Aug 8, 2011… A single mosquito bite can give you West Nile virus. ... Many people who are bitten by an infected mosquito won't get sick—many others aren't ..."*

The resulting snippet phrases were then:

- "Aug 8, 2011",
- "A single mosquito bite can give you West Nile virus." and
- "Many people who are bitten by an infected mosquito won't get sick—many others aren't".

Each snippet phrase was then stored along with its associated URL for the top URLs (the number of URLs used varied, as explained below). These snippet phrases were used to create a cluster of words semantically related to the original search terms in the following manner: I downloaded the web page corresponding to each of the snippet phrases and stripped out the HTML, php, and javascript code. Then the snippet phrase was located in the web page. If the snippet phrase was located in the meta data (e.g., the title, the page description, any tag within the head section, etc.), the nonstop words in the meta data were added to the cluster of words. If the snippet phrase was located in the web page itself, the 50 nonstop words before the snippet and the 50 nonstop

words after the snippet phrase were added to the cluster of words associated with the original search terms. The count of the occurrence of each of the words in the cluster was recorded and updated by the number of times it occurred in the surrounding text. This was done for each of the top snippet phrases.

Originally I used all snippet phrases associated with the top 20 URLs returned from the Google search. However, if you look at the snippet phrases associated with the West Nile Virus question (listed previously), it is obvious that not all snippet phrases are directly related to the search, and some snippet phrases seem to be more related to the search than others. For example, if available, Google will include an article date (e.g., "Aug 8, 2011") because it helps searchers evaluate whether or not to read a page, but it is unrelated to the search terms.

I thus decided to modify my original approach of taking all snippet phrases associated with the top 20 URLs. Instead, I take the top 50 URLs (instead of 20), and then use only the snippet phrases with the most search terms (i.e., the nonstop words from the original question). I decided to use only snippet phrases with the most search terms for the following reason: While the entire snippet might contain all the search terms from the question, when a snippet was broken down to snippet phrases, a snippet phrase might contain only one of the search terms. Thus the text surrounding that particular phrase is unlikely to be related to the question. The overarching goal is to locate text relevant to the entire question within a document that is on one or a limited set of topics. If, for instance, a snippet phrase contained only one of the search terms, the search term contained does not tell us much about the topic of the surrounding text in the document and thus does not tell us about its relevance to the topic of the question.

Along with using only snippet phrases that contained all or most of the question terms, I also increased the number of URL snippets used to 50. I did this because certain topics are significantly more commonplace than other topics, and thus discussed more frequently in relation to the topic of the entire question. If the question's topic is discussed more frequently, it is likely that a wider array of words are used in discussion of the question's topic. By expanding the number of URLs used, it was hoped that there would be more snippet phrases that contained all the search terms from the question, and thus the cluster formed from the text surrounding the numerous snippet phrases with all the search terms would contain more of the various topics discussed in relation to all the search terms. So, for instance, a question on how marijuana affects the brain could return many snippet phrases that contain all or most of the search terms from the question. By using many snippet phrases from more URLs, I wanted to get subtle and less common words associated with this question in the resulting cluster as well as the most common ones. On the other hand, for less common topics the expanded set of URLs is unlikely to bring in noise because I am only taking the snippet phrases that have all or most of the question's search terms in them. As a result, the less commonly discussed questions will result in a smaller cluster of words resulting from text surrounding fewer snippet phrases.

4.2.2.1 Creating the Word Clusters

Specifically, to modify the system to use the snippet phrases with the most search terms, I first increased the number of URLs used to about 50[5]. I took the snippet associated with each of the URLs, separated it into snippet phrases as

[5] The exact number of URLs used for this part varied because some of the Web pages associated with the returned URLs were either no longer valid or the text within the Web page had been modified.

described above, and stored each snippet phrase with its accompanying URL. The total set of snippet phrases were then ordered by the number of search terms that occurred within that snippet phrase. Those snippet phrases with the most search terms were then used in the formation of the cluster, and the other snippet phrases were discarded. For instance, if there were 6 search words used in the Google search that returned the list of URLs and their associated snippets, each of the snippet phrases created from this list of snippets contained between 1 and 6 of the search words. In this example I first used all snippet phrases (and their associated URLs) with at least 5 of the search words to create a word cluster, unless there were less than 4 snippet phrases with 5 or more search words, in which case I included all snippet phrases with at least 4 search words, etc. All other snippet phrases with fewer search terms were discarded. Thus only the snippet phrases with the most search terms were used to locate text within a Web page, and only the text surrounding the snippet phrases with the most search terms were used to form the word cluster.

4.2.2.2 Resulting Word Clusters

The entire process resulted in clusters of words with a loose semantic connection to the search terms, or the question. Table 3 and Table 4, below, give examples of the top words in a cluster created using this method.

The resulting clusters are clearly related to the search terms, or question. They also exhibit the loose semantic connection of things frequently discussed together (e.g., in Table 3, below, the connection between ‘virus’ and ‘cases’, ‘risk’, etc.). Interestingly, word disambiguation seems to be occurring automatically. For example, in Table 3, the word ‘catch’ is automatically disambiguated in this cluster. Top words in the cluster include, ‘infected’, ‘health’, ‘disease’, ‘symptoms’, ‘illness’, ’fever’,

'infection', 'risk', etc., all of which could be associated with the meaning of 'catch' that is associated with becoming infected with a disease. There are no words in the cluster that are associated with other meanings of the word 'catch', e.g., 'mitt', 'ball', 'game', etc.

Table 3 **Results from Question:** How do people catch the West Nile Virus?
Query Terms: *'people', 'catch'. 'west', 'nile', 'virus'*
Most frequently occurring words in resulting cluster and their counts:

virus: 237	water: 25	time: 14	white: 10	avoid: 8
west: 226	chicago: 23	develop: 14	boston: 10	larvicide: 8
nile: 200	mild: 22	services: 14	residents: 10	middlesex: 8
mosquito: 132	melrose: 21	bites: 13	bitten: 10	body: 8
people: 105	get: 20	cause: 13	pool: 10	dawn: 8
mosquitoes: 99	pools: 20	1: 13	population: 10	commonly: 8
health: 85	take: 20	culex: 13	larvae: 10	story: 8
catch: 82	bird: 20	species: 13	use: 10	way: 8
infected: 76	severe: 20	news: 13	one: 10	wayland: 8
can: 65	local: 19	meningitis: 13	jul: 10	posted: 8
will: 48	2012: 18	like: 13	transmitted: 9	neighborhoods: 8
disease: 47	also: 18	become: 13	percent: 9	test: 8
wnv: 45	include: 18	reduce: 12	confirmed: 9	carried: 8
control: 41	common: 18	county: 12	breeding: 9	animals: 8
symptoms: 41	human: 18	measures: 12	eee: 9	prevent: 8
said: 39	may: 17	july: 12	last: 9	standing: 8
birds: 37	treatment: 17	state: 11	eastern: 9	earlier: 8
cases: 37	precautions: 17	prevention: 11	according: 9	apply: 8
basins: 37	humans: 17	area: 11	officials: 9	long: 8
positive: 34	first: 17	many: 11	dusk: 9	several: 8
illness: 32	roxbury: 16	dead: 11	identified: 9	commission: 8
fever: 32	detected: 16	traps: 11	number: 9	following: 8
city: 30	sick: 16	50: 11	home: 9	days: 8
public: 30	bite: 16	protect: 11	unit: 9	name: 8
infection: 29	new: 16	staten: 11	paint: 9	roslindale: 8
encephalitis: 29	board: 15	wicked: 11	reported: 9	surveillance: 8
department: 29	spread: 15	testing: 11	service: 9	site: 8
found: 27	headache: 14	island: 11	around: 9	photos: 8
risk: 26	serious: 14	type: 10	high: 9	borne: 8
year: 26	areas: 14	treated: 10	smith: 8	
basin: 25	east: 14	pm: 10	share: 8	

Table 4 **Results from Question:** What dietary factors are thought to raise and lower cholesterol?
Query Terms: '*dietary', 'factors', 'thought'. 'raise', 'lower', 'cholesterol'*
Most frequently occurring words in resulting cluster and their counts:

cholesterol: 116	disease: 13	may: 7	attack: 5	includes: 4
eggs: 44	also: 12	increase: 7	u: 5	americans: 4
levels: 33	healthy: 12	coronary: 7	diseases: 5	fatty: 4
dietary: 26	raise: 12	might: 7	moderate: 5	stroke: 4
saturated: 26	hdl: 11	two: 7	al: 5	several: 4
high: 24	thought: 11	consuming: 7	et: 5	tend: 4
fats: 23	risk: 10	myplate: 6	calories: 5	contains: 4
affect: 22	mediterranean: 10	exercise: 6	diabetes: 5	women: 4
diet: 22	one: 10	low: 6	three: 5	carbohydrates: 4
heart: 21	food: 9	mg: 6	age: 5	body: 4
can: 17	eating: 9	see: 6	studies: 5	bad: 4
egg: 17	2: 9	level: 6	many: 4	genetics: 4
fat: 16	weight: 9	smoking: 6	reduce: 4	fitness: 4
blood: 16	well: 9	whites: 5	cause: 4	use: 4
factors: 16	ldl: 8	percent: 5	people: 4	acids: 4
lower: 15	1: 8	source: 5	family: 4	2009: 4
person: 14	b: 8	good: 5	things: 4	extent: 4
foods: 14	eat: 8	amount: 5	triglycerides: 4	keys: 4
health: 13	day: 8	related: 5	less: 4	us: 4
total: 13	atherosclerosis: 7	varies: 5	makes: 4	recent: 4
				however: 4

4.2.3 Adding Global Meaning Weight

The clusters created so far clearly exhibited the loose semantic relationship I was hoping to capture. However, certain words occurred frequently that contribute little to the semantic connections of the cluster, whereas certain words that occurred infrequently in a cluster had a much greater semantic link to the overall meaning of the cluster. For example, in the word clusters in Table 3 and Table 4, the word 'can' occurred frequently in both clusters, occurring 10th most frequently in the first cluster and 11th most frequently in the second cluster. Yet the word 'can' contributes little in terms of semantic meaning in the cluster. Indeed, it is a very common word that will most likely occur frequently in almost every cluster, regardless of the search terms used. The same can be said for words like 'also' and 'may'. In contrast, the word

'neurotransmitter', for example, occurred only 3 times in the cluster associated with the question 'How does marijuana affect the brain?', yet this word most likely rarely occurs in clusters on other topics and thus most likely has more significance in terms of the semantic meaning of the cluster. I wanted to make sure that words with little semantic meaning had significantly less weight in determining the connection between the cluster and the paragraph text than those words with more semantic meaning. I wanted to use something akin to TF-IDF, but calculating the IDF part for documents in the World Wide Web posed a problem. I chose to calculate a "Global Meaning Weight" (explained below) which is an Inverse Document Frequency (IDF) precalculated for a large set of words that may be encountered during the snippet processing.

4.2.3.1 Global Meaning Weight Calculations

Because the system uses the World Wide Web to create the clusters, I wanted to make sure that the Global Meaning weight (GM weight) reflected the relative occurrence of words in Web pages throughout the WWW. For this weight, my goal was not to accurately weight every word on the Web but to make sure words that occurred frequently on the Web held a lower weight than those that occurred infrequently. In TF-IDF, the IDF factor is the factor that determines the general relevance of a particular term. In general it is defined as the total number of documents in a corpus divided by the total number of those documents a particular word occurs in. For my purposes I wanted to get a general idea of how often words were likely to occur in a wide range of Web page. Ideally my corpus would be all the Web pages in the World Wide Web, and I'd have a count of how many Web pages every word in the World Wide Web occurred in (i.e., a table of every word on the

World Wide Web and the count of the number of Web pages that word occurred on). Clearly this is impractical, and, for my purposes, unnecessary. Instead I decided to take a relatively random subset of Web pages (which acted as my corpus of documents in IDF terminology), and created a table of the words that occurred in that Web page subset as well as the count of the number of Web pages in that corpus that the word occurred in.

The table of Global Meaning weights was calculated by first generating approximately 2260 random word pairs (from "erudite quixotic" to "tissue calendar"). For each of these word pairs I did a Web search using Google. The top 4 URLs for each word pair became the corpus, or random set of Web pages used in calculating the GM weights. For each of the approximately 4 top URLs associated with each word pair, I took the snippet associated with it, divided the snippet into phrases separated by "…" as described in the section on My Approach to Identifying Relevant Words. The snippet phrases were located either in the content of the Web page or its meta tags, and each unique word in the 50 nonstop words above and below or the meta tag surrounding the snippet phrase was added to the GM weight table, with counts of words already in the table increased by 1. The result was a set of 21,429 random words and the count of the number of documents the words occurred in. The table of GM weight, or IDF weights, was then calculated as follows: For each word Wq in the table, I calculated its weight using the formula:

$$\text{globalGM}(W_q) = \log\left(\frac{D}{CDW_q+1}\right)$$

Where

D = total number of documents (Web pages) used to calculate the GM values (e.g., the size of my corpus of documents), and

CDWq= count of documents in corpus that contained the word Wq one or more times.

Because the word set is relatively small (compared to all the words in existence), many words exist that are not part of this word set. I used add-one smoothing to compensate for the sparse data by adding one to each of the word counts.

The GM weight table was calculated ahead of time because it was relatively time-consuming. My system needs to run in real time for users to find it useful, and calculating the GM weight each time would slow down the system to the point of making it notably less useful. In addition, there was no need to run it every time the system seeks to match a question and paragraph texts. The approximate counts of word occurrence in the World Wide Web is not something that changes drastically on a daily basis. Thus to keep the system running as quickly as possible without sacrificing the use of Global Meaning weights, I created the table ahead of time and expect to recreate the table approximately every 6 months.

Once the table of GM weight was created, I wanted to use it with the word clusters formed from the question's search terms so that the weight of the words in the cluster would reflect their global significance as well as their relevance within the cluster itself. For this I used the GM weight as the IDF factor in TF-IDF calculations, and for the TF factor for a particular question's word cluster, I used the count of a word CWi in the cluster formed based on the question, normalized by the total number of words in the question cluster CWT. Thus, for each word instance in a question cluster, its weight was calculated as follows:

$$\text{Global\_TF\_IDFW}_i = (\text{CW}_i/\text{CWT}) * \text{GMW}_i$$

Where:

CW_i is the count of word Wi in a question cluster (i.e., the total number of times the word occurred in text surrounding the snippet phrases)

CWT is the total count of all words in a question cluster (e.g., if a cluster consisted only of "big", occurring 5 times, and "balloon", occurring 3 times, the CWT would be 8)

GMW_i is the Global Meaning weight of word W_i, calculated as described above

GlobalTFIDF weight-ordered cluster are shown in Table 5 and Table 6, below.

Table 5 **Results from Question:** How do people catch the West Nile Virus?
Query Terms: *'people', 'catch'. 'west', 'nile', 'virus'*
Resulting cluster ordered by Global_TFIDF weight multiplied by 100

nile: 25.9	department: 1.8	higgs: 1.1	boston: 0.8	get: 0.7
virus: 22.6	severe: 1.8	develop: 1.0	swollen: 0.8	percent: 0.7
mosquito: 16.3	chicago: 1.7	prevention: 1.0	confirmed: 0.8	reported: 0.7
west: 14.3	headache: 1.7	include: 1.0	cause: 0.8	carried: 0.7
mosquitoes: 13.8	public: 1.6	local: 1.0	island: 0.8	commission: 0.7
wnv: 7.9	can: 1.6	dusk: 1.0	dead: 0.8	newton: 0.6
infected: 7.8	detected: 1.6	measures: 1.0	identified: 0.8	unit: 0.6
catch: 6.4	city: 1.6	serious: 1.0	rash: 0.8	50: 0.6
encephalitis: 5.1	staten: 1.5	common: 1.0	population: 0.7	coma: 0.6
basins: 4.6	bird: 1.5	species: 1.0	paint: 0.7	standing: 0.6
health: 4.4	larvicide: 1.4	transmitted: 1.0	lymph: 0.7	smith: 0.6
people: 3.9	wayland: 1.4	board: 1.0	fresno: 0.7	tested: 0.6
melrose: 3.7	bite: 1.4	surveillance: 0.9	transmit: 0.7	commonly: 0.6
symptoms: 3.4	humans: 1.4	east: 0.9	officials: 0.7	may: 0.6
disease: 3.4	water: 1.4	borne: 0.9	dawn: 0.7	carrying: 0.6
fever: 3.3	sick: 1.3	take: 0.9	services: 0.7	first: 0.6
illness: 2.9	bites: 1.3	reduce: 0.9	july: 0.7	also: 0.6
birds: 2.8	found: 1.3	areas: 0.9	eastern: 0.7	pm: 0.6
roxbury: 2.8	traps: 1.3	breeding: 0.9	become: 0.7	bradford: 0.6
infection: 2.7	eee: 1.3	convulsions: 0.9	ameara: 0.7	area: 0.6
cases: 2.6	will: 1.3	gwillimbury: 0.9	stein: 0.7	animals: 0.6
positive: 2.5	treatment: 1.2	testing: 0.9	cdph: 0.7	earlier: 0.6
control: 2.5	larvae: 1.2	glands: 0.9	efrat: 0.7	treating: 0.6
basin: 2.4	bitten: 1.1	2012: 0.9	3617: 0.7	hyde: 0.6
culex: 2.3	spread: 1.1	protect: 0.8	simcoe: 0.7	biting: 0.6
meningitis: 2.3	middlesex: 1.1	treated: 0.8	boson: 0.7	outdoor: 0.6
mild: 2.0	roslindale: 1.1	jul: 0.8	larvicides: 0.7	avoid: 0.6
pools: 2.0	human: 1.1	neighborhoods: 0.8	spraying: 0.7	prevent: 0.6
precautions: 1.9	wicked: 1.1	residents: 0.8	disorientation: 0.7	white: 0.6
said: 1.8	year: 1.1	county: 0.8	crows: 0.7	
risk: 1.8	aches: 1.1	pool: 0.8	clovis: 0.7	

Table 6 **Results from Question:** What dietary factors are thought to raise and lower cholesterol?
Query Terms: '*dietary', 'factors', 'thought'. 'raise', 'lower', 'cholesterol'*
Resulting cluster ordered by Global_TFIDF weight multiplied by 100

cholesterol: 41.4	raise: 3.1	imagespolka: 1.7	also: 1.3	dares: 1.1
eggs: 11.9	disease: 3.1	myquit: 1.7	polka: 1.3	t7gwlcjgyfk: 1.1
saturated: 9.6	total: 2.9	stents: 1.7	stroke: 1.3	keys: 1.1
dietary: 9.5	healthy: 2.8	calories: 1.6	might: 1.2	low: 1.1
fats: 8.7	coronary: 2.7	increase: 1.6	al: 1.2	recommends: 1.1
levels: 7.8	myplate: 2.7	moderate: 1.6	percent: 1.2	day: 1.1
hdl: 6.3	person: 2.6	diabetes: 1.6	yolk: 1.2	fitness: 1.1
diet: 5.9	triglycerides: 2.3	food: 1.6	well: 1.2	amount: 1.1
affect: 5.8	risk: 2.3	fatty: 1.6	attack: 1.2	americans: 1.1
ldl: 4.6	mg: 2.2	varies: 1.5	extent: 1.2	tend: 1.0
egg: 4.5	eating: 2.2	exercise: 1.5	studies: 1.2	age: 1.0
heart: 4.2	health: 2.2	b: 1.5	cad: 1.2	intake: 1.0
factors: 4.0	consuming: 2.2	diseases: 1.4	level: 1.2	reduce: 1.0
fat: 3.9	thought: 2.1	carbohydrates: 1.4	qefzz1: 1.1	quit: 1.0
blood: 3.7	eat: 1.9	can: 1.4	lwdons: 1.1	obesity: 0.9
foods: 3.7	whites: 1.9	shannan: 1.4	1p: 1.1	source: 0.9
high: 3.6	weight: 1.9	yolks: 1.4	stent: 1.1	contains: 0.9
lower: 3.4	smoking: 1.8	acids: 1.4	hwrf: 1.1	2: 0.9
mediterranean: 3.3	monson: 1.7	genetics: 1.3	jjdigilio: 1.1	balanced: 0.9
atherosclerosis: 3.2	dotgetty: 1.7	et: 1.3	ancel: 1.1	bray: 0.9
				unstoppable: 0.9

Intuitively the top words in the cluster are more related to the question's topic.

4.3 Matching Cluster Words to Paragraphs

Once I had created these clusters of words most likely to be semantically related to the question, I now needed to establish a method for matching the cluster with the appropriate paragraph text in a document associated with the question responsible for the cluster.

4.3.1 Baseline Matching of Cluster Words to Paragraphs

For this method, I used the same technique used in the Baseline method, only the search terms were expanded to include all cluster terms and weights. The

relevance of a particular paragraph to a question was determined as follows. Given a document containing *P* paragraphs $P_1 \ldots P_p$ and an expanded set of *n* search terms in the Cluster of Words created from the Google search using the nonstop words in the associated question, $W_1 \ldots W_n$, I scored each paragraph as follows:

$$\text{Score}(P_j) = \sum_{i=1}^{n} \left(\frac{\text{CW}_i}{CWP_j}\right) * \log\left(\frac{P}{CPW_i}\right) * Global\_TFIDFWi$$

Wh

4.3.2 Results of Baseline Paragraph/Word Cluster Matching

The results of using this method to rank paragraphs is shown in Table 7, below.

Table 7 Rankings of the paragraphs using both Document- and Global- TF-IDF weights combined

Question/ Document	Ranking of paragraph ranked highest by visual scanners	Ranking of paragraph holding the answer	Total # paragraphs in Document
QD1-La	7	7	14
QD2-St	8	8	15
QD3-Wn	2/5	2	12
QD4-Co	6	11	14
QD5-Ci	2	6	11
QD6-Ea	4	4	13
QD7-Mo	6	6	10
QD8-Pi	8	8	15
QD9-Ma	2	2	12
QD10-Me	13	8	25

4.3.3 Results Discussion

I was surprised and somewhat disappointed in the results. The clusters appeared to exhibit the loose semantic relationships I was hoping to capture. Yet using the clusters to identify paragraphs identified by scanners as relevant to the question's topic resulted in only marginally better results (for three of the ten documents, the paragraph most focused on was the second-highest ranked) over previous methods. Upon further examination, however, I noticed that, while the words in the cluster that are weighted highest were very relevant, there were many words in the cluster, and a number of the words had very low weights, meaning they both occurred infrequently in the cluster combined with a very low Global Meaning Weight. For example, in the cluster created from the search terms *'people', 'catch'. 'west', 'nile', 'virus'*, the entire cluster contained 1659 unique words. Of those words, 732 occurred only once, and another 375 occurred only twice. For the cluster created from the search terms

‘*dietary’, ‘factors’, ‘thought’. ‘raise’, ‘lower’, ‘cholesterol’* (one of the smallest clusters), there were 828 words in the cluster. Of those, 493 occurred only once and another 159 occurred only twice. Many of these words had very low Global Meaning weights as well. Words that occurred only infrequently probably contributed little to the semantic relatedness of the cluster and may have occurred by chance, especially if they had a low Global Meaning weight. In fact, the larger the word cluster, the more likely it is that words occurring infrequently and with low Global Meaning weight contribute much when using the clusters to identify relevant paragraph text, and in fact possibly confound the results because these infrequent words are in essence “noise”. Thus I decided to only use the top 25% of the cluster words, based on their Global TF-IDF weighted value.

4.4 Most Relevant Information Ranking (MRI)

To eliminate “noise” words, I ordered the cluster words by their Global-TFIDF weight, and eliminated the bottom 75%, thus using only 25% of the cluster words with the highest weight to rank the paragraphs in the document.

On examining the paragraphs, it appeared that many longer paragraphs had significant content that wasn’t necessarily related to the question’s topic. Indeed, it appeared when looking at scanner’s data, that many focused on a particular part of the paragraph, or on certain parts of the paragraph and focused less on others. Thus it seemed that a more accurate way of ranking a paragraph would be to rank it based only on the most relevant text in the paragraph, again eliminating a lot of “noise” that might confound the results.

To accomplish this, the method I used was to give each sentence in the document a rating score (based on the matching method described in Baseline

Matching of Cluster Words to Paragraphs, above, using only the top 25% of the related word cluster), then use the top 25% most relevant sentences to rank paragraphs. Specifically, I took each sentence in the entire document and gave it a relevance score by matching the words in the reduced cluster to the words in the sentence, weighted by multiplying its Document TF-IDF weight and its Global TF-IDF weight. The sentences were then ordered based on their relevance score. Because sentences with lower relevance scores provide little useful information, I used only the sentences with the top 25% relevance score. Those sentences were ordered, and given a number based on the inverse of their ranking. Then each paragraph's relevance score was calculated by adding the number assigned to each sentence belonging to that paragraph.

So, as an example, in a document with 37 sentences, each sentence is given a relevance score. The top 9 are saved. The sentence with the highest relevance score is given a 9, the sentence with the second highest would receive an 8, etc. A paragraph's relevance score is calculated based on the sum of the scores of its sentences. So if a paragraph holds sentences ranked 1, 2, and 7, it would be given a relevance score of 9+8+3, or 20.

All paragraphs were given a relevance score in this manner, and then ordered according to rank.

4.4.1 MRI Results

Results of using the Most Relevant Information (MRI) method to rank the paragraphs are shown in Table 8, below.

Table 8 Rankings of the paragraphs using the MRI method

Question/ Document	MRI ranking of paragraph ranked highest by visual scanners	MRI ranking of paragraph holding the answer	Total # paragraphs in Document
QD1-La	1	1	14
QD2-St	1	1	15
QD3-Wn	1/2*	1	12
QD4-Co	1	3	14
QD5-Ci	1	4	11
QD6-Ea	4	4	13
QD7-Mo	5	5	10
QD8-Pi	6	6	15
QD9-Ma	7	7	12
QD10-Me	10	Unranked**	25

**For QD3, 2 paragraphs tied as the most focused on. My method ranked those 2 paragraphs as the first and second most relevant.*
***If a paragraph contained no sentences with a top 25% ranking, it was unranked.*

Results show that for 5 of the 10 questions, the paragraph ranked most relevant was the one most focused on during the scanning experiments (see Table 1 column 3). This is noteworthy because it suggests that the MRI method is making the semantic connections in a manner similar to how visual scanners make connections. Indeed, in 3 of the questions, the paragraph most focused on by visual scanners was not the paragraph with the answer, and for 2 of these three questions the MRI method identified the paragraph focused on and not the paragraph with the answer as the most relevant paragraph. In looking at the sentences identified as most relevant, indeed there appears intuitively to be a strong connection between the question and the sentences identified as most relevant. For instance, in the question,

"How do people catch the West Nile Virus?"

the sentences identified by the MRI method as the most relevant include (in order) (punctuation has been stripped):

1. west nile virus
2. it is spread by mosquitoes
3. transmission happens when a mosquito bites a bird infected with the west nile virus and the virus enters the mosquito bloodstream
4. most people infected with the west nile virus have no signs or symptoms
5. most people recover from west nile virus without treatment
6. to help control west nile virus eliminate standing water in your yard
7. about 20 percent of people develop a mild infection called west nile fever
8. some laboratory workers involved in west nile research have contracted the disease from infected animals
9. mosquitoes breed in pools of standing water
10. in rare cases it is possible for west nile virus to spread through other routes including
11. watch for sick or dying birds and report them to your local health department
12. west nile virus is common in areas such as africa west asia and the middle east
13. in the united states wild birds especially crows and jays are the main reservoir of west nile virus but the virus is actually spread by certain species of mosquitoes
14. your best bet for preventing the virus and other mosquito borne illnesses is to avoid exposure to mosquitoes and eliminate mosquito breeding sites
15. your overall risk of contracting west nile virus depends on these factors time of year
16. then the infected mosquito bites an animal or a human and the virus enters the host bloodstream where it may cause serious illness
17. even if you are infected your risk of developing a serious west nile virus related illness is extremely small

4.4.2 nDCG Comparison of MRI and Other Methods

In analyzing the effectiveness of the MRI method in identifying most relevant paragraphs, I wanted to see not just how well it identified the paragraph most focused on in the visual scanning studies, but how well it accurately identified both relevant and less relevant paragraphs as determined by the visual scanning studies. Visual scanners appear to focus more heavily on area in the document with text related to the question, and less heavily on area with less relevance. My method should also identify areas more and less semantically related to the question.

To test this, I used the Discounted Cumulative Gain measure, normalized (nDCG) (Jarvelin and Kekalainen, 2002). The nDCG is used to measure how well one ranking compares to another ranking, assuming the importance of the items appearing

higher in the ranking is higher than the importance of items appearing lower in the ranking. It is often used in determining the effectiveness of the Web page URLs returned from a Web search by determining how effective the ranking is in placing most relevant documents before documents of lesser relevance. In those cases a document's relevancy is judged on a 0 to 3 scale where 0 is not at all relevant and 3 is very relevant. To mimic this, I used the ordering of the paragraphs as determined by the focus studies as the ideal case. I took the number of paragraphs in the document, and divided by 4. The top 25% were assigned an importance score of 3, the next 25% were assigned a 2, the next 25% were assigned a 1, and the bottom 25% were given an importance score of 0. I then calculated the nDCG for the rankings from the Baseline Method, the Synonyms, Hypernyms, and Hyponyms (SHH) Method, the Documeng and Global TF-IDF Method, and the MRI Method. Results are shown in Table 9, below.

Table 9 nDCG scores for each of the documents using the different methods for calculating ranking

	Baseline	SHH	TFIDF	MRI
QD1-LA	0.842	0.819	0.743	0.853
QD2-St	0.824	0.676	0.83	0.819
QD3-Wn	0.747	0.576	0.799	0.877
QD4-Co	0.789	0.789	0.787	0.885
QD5-Ci	0.858	0.73	0.854	0.83
QD6-Ea	0.896	0.931	0.95	0.981
QD7-Mo	0.573	0.784	0.82	0.862
QD8-Pi	0.732	0.746	0.64	0.813
QD9-Ma	0.498	0.841	0.814	0.824
QD1-Me	0.732	0.688	0.702	0.679
Average:	0.7491	0.758	0.7939	0.8423

The MRI method is statistically significantly better than the Baseline Method and the synonym/hypernym/hyponym method. While the results trended better than the TF/IDF method, results were not statistically significantly better than the TF/IDF method.

On average, the MRI method did notably better than any of the other methods. This is important because it indicates that the MRI method is better than the other methods at identifying not just the most focused on paragraph, but also the paragraphs most focused on by visual scanners.

Indeed, this is impressive because the MRI method did have some disadvantages. In using only the top 25% of sentences to rank paragraphs, the MRI method didn't give any ranking to a certain number of paragraphs. For those paragraphs with no ranking, I gave them all an equal ranking of the last place. Because the nDCG method devalues those paragraphs at the bottom of the ranking, all those last-place paragraphs could only receive a small value at best, and thus could never add a lot of value to the overall score. This hurt the MRI method because visual scanners apparently combined both semantic relationships and visual cues when determining which paragraphs they focused on, and thus often focused on the first paragraph in a document, resulting in this first paragraph having a high ranking (often second or third). The MRI method identified paragraphs solely based on semantic relationships. In not taking these visual cues into account, the MRI method would rank this first paragraph quite poorly, thus adversely affecting its nDCG score.

4.4.3 Testing Word Cluster Size

Once I had established that the MRI method was in fact doing a good job of ranking the paragraphs relevant to those focused on by visual scanners, I wanted to see if the size of the cluster of words used in the MRI method to rank the paragraphs would affect its effectiveness. Recall that I used the top 25% of the words in the semantically-related word cluster for determining a sentence's score. I wanted to see if different sized clusters made the overall paragraph rankings better or worse. I thus reranked everything with a cluster size of of 5%, 15%, and 35% as well as 25%. The results are shown in Table 10, below.

Table 10 nDCG scores of MRI run with different size word clusters

	MRI (05%)	MRI (15%)	MRI (25%)	MRI (35%)
QD1-La	0.876	0.878	0.853	0.868
QD2-St	0.857	0.865	0.819	0.811
QD3-Wn	0.877	0.874	0.877	0.867
QD4-Co	0.823	0.851	0.885	0.823
QD5-Ci	0.832	0.861	0.83	0.861
QD6-Ea	0.978	0.978	0.981	0.99
QD7-Mo	0.837	0.841	0.862	0.841
QD8-Pi	0.713	0.699	0.813	0.685
QD9-Ma	0.824	0.824	0.824	0.824
QD10-Me	0.646	0.679	0.679	0.679
Average:	0.8263	0.835	0.8423	0.8249

The rankings generated by the MRI method using the top 25% of the word cluster resulted in the best rankings.

4.4.4 Limitations of Current MRI method

It must be mentioned that the MRI method was developed using only 10 text documents and their related questions. Ideally I would have liked to develop the MRI method for identifying related text within a document, and then test it on a separate set of documents (on which I'd collected additional visual scanning information). However, the MRI method used a Google app that allowed me to do in-program web searches and garner the resulting search information in a text document. That app was deprecated shortly after the collection of data for the 10 documents used in this experiment. Thus I was forced to limit my test data to the original 10 documents.

It also must be noted that the research conducted for the process of finding semantic connections similar to those made by people scanning through documents was completed over five years ago. Since then notable advances have been made in the field of semantic relationships. The most notable of these are word2vec and GloVe. word2vec takes as input a text corpus and produces a set of co-occurrence vectors that are fed into a neural net. The result is a vector space with word vectors with similar context occurring close together in the vector space. This method was developed by researchers at Google, and can be used to either predict a word based on surrounding words, or it can be used to predict surrounding words based on a word (Mikolov et al., 2013). Global vectors for word representation, or GloVe, was developed by researchers at Stanford University and is a count based model, meaning that it uses matrices of co-occurrence information (i.e., how often "leash" occurs within x words of "dog"). The probability of how closely two words are related is calculated as the probability of the two words in the co-occurrence matrix divided by the maximum probability of two words co-occurring (Pennington et al., 2014). Both methods have drawbacks: word2vec requires a vast corpora of data in order to make

the vector space and GloVe constructs an in-memory matrix that requires a great deal of RAM. However, both of these methods should be evaluated in terms of their potential to improve the identification of areas of text semantically related to a question.

4.4.5 Summary of Work So Far

At this point it has been established that visual readers do scan for semantically related text within a document when scanning for the answer to a complex question. I have also shown that it is possible to replicate those semantic connections in a reasonable amount of time across contexts (i.e., in an open domain). The next step was to develop a user interface to allow users of assistive technology to find the semantically related text.

Chapter 5

USER INTERFACE

The next step in creating a functional system for conveying information was to create a user interface. At its core, the system needed to lead users to information within the document related to the question and, once there, allow the user to switch out to normal reading mode if they so desired. The system needed to convey information about the overall topology, and to convey information about where within the document large quantities of information related to the question were located. In addition, in order for the system to be adopted and accepted, there were specific requirements related to the interface itself. The interface needed to be simple and easy to use in terms of both entering the question(s) and the accompanying document, and in utilizing the system's output. It needed to keep the document intact, so users could read the entire document if they wanted to, yet it needed to allow users to move throughout the document as if they were scanning through it. It needed to be easily navigable, including the ability to allow the user to move forwards and backwards within the document (similar to how visual scanners focus on something relevant, then slow down and read the text surrounding the area they focused on). It had to allow users to easily and quickly change the way in which they were scanning through the document (e.g., to go from reading to scanning, to go from a topological scan to a scan purely based on relevance to the question, etc.) It had to either utilize a system users were familiar with, or be a new system that users could learn to use quickly and intuitively. The data used in updating the interface needed to be generated relatively

quickly and to allow for almost instantaneous navigation. And, of course, it had to work with the most commonly used reading assistive technologies. Towards this end, I decided to create a dynamically generated web page out of the document, including links to areas within the document based on the question and links within the document to the next area of interest. The Web page allows users to switch modes of navigation simply by switching keystrokes. Users can also switch modes of navigation via the links generated and placed at the top and within the page simply by clicking on a button signifying the mode of navigation they wish to use. Thus within each web page, users can switch modes of navigation either by clicking on a button that generates a new web page with links at the top of the page based on the user's choice of navigation mode, or at any time simply by switching the keystrokes she or he is using.

5.1 User Interface Research

A good deal of research has been done on both how people who are blind navigate through web pages and on ways in which to improve a person who is blind's experience of navigating through web pages.

5.1.1 Web Page Accessibility

Web pages today are largely written in HyperText Mark-up Language (HTML), a mark-up language that includes a set of tags labeling each part of the web page. The tags are then rendered by a browser to look like the web pages we see. Web page tags have default styles associated with them (e.g., the <p> tag, or the paragraph tag is usually rendered by most browsers as black text on a white background with Times New Roman font). Those styles can be overridden using Cascading Style

Sheets (CSS), which define styles for a particular tag or set of tags. While HTML and CSS are becoming more intermingled in more recent versions, in general HTML allows the author to define what an element on a web page is, and CSS allows the author to define how that element should be displayed on a web page. CSS includes font color, background color, background images, line spacing, indents, font size, and much more. HTML and CSS are used to create static web pages. However, today many web pages have added elements that allow them to change dynamically based on input from the user of the web page. A simple example would be rolling your mouse over a smaller image and having a larger version of that image appear in place of another image, as well as text appearing in a paragraph describing that image. Some dynamic web pages use client-side scripting languages (meaning the code is rendered by the browser), like javaScript, whereas others uses server-side scripts (run and rendered on a web server, and sent to the browser) written in PHP, Perl, or a number of other languages.

While static web pages have their own set of accessibility issues (e.g., contrast between background and foreground text, alternative text for images, etc.), dynamically changing web pages pose their own set of challenges for those using assistive technology. In creating a dynamic web page, both sets of issues must be addressed as thoroughly as possible.

5.1.1.1 World Access Initiative (WAI):

In order to address issues of web page accessibility for static web pages, the Web Access Initiative (WAI), launched by the World Wide Web Consortium (W3C) has a set of guidelines (Web Content Accessibility Guidelines, or WCAG 2.0) for designing web pages. The guidelines guide web page authors on how to create web

pages that are accessible for not just people who are blind and low vision, but for people with a wide range of varying abilities including people with cognitive limitations, limited motor function, photosensitivity, etc. While the visual guidelines include those for making images more accessible, layout modification for easier navigation, guidelines for blinking or moving content that might induce seizures, etc., also included are basic guidelines for making sure that a web page is usable without the use of a mouse (i.e., all functionality can be accessed via the keyboard), that the text is readable for both users who are low-vision, and color-blind, and that text is readable through a screen reader. It also includes basic guidelines to make sure that a web page is accessible using different assistive technologies (*Web Content Accessibility Guideline, 2011*).

5.1.1.2 Accessible Rich Internet Applications (ARIA)

The WCAG guidelines consist largely of technology-neutral guidelines for web pages (e.g., all non-text items must have an alternative text version, web page text should be resizable, etc.). In general they do not specify exactly how everything should be implemented. In addition, they do not include intelligent decisions about understanding, labeling, and dealing with content on a web page, especially dynamically updating content (e.g., widgets, drop-down menus, changing content, etc.) For this, the WAI created the Accessible Rich Internet Applications (ARIA). ARIA are a set of specifications designed to make user interfaces and dynamically changing content of web pages more accessible. ARIA are attributes (tags, labels, etc.) added by the web page author that standardize how to implement some of the more challenging guidelines put forth by the WCAG. With ARIA, web page authors are able to define the 'roles' of areas within the web page (e.g., 'navigation' for a

collection of links for navigating the site, ‘menu’ for a widget that offers the user a list of choices, etc.), the ‘states’ of those areas (e.g., ‘checked’ for when a checkbox or radio button in a form is checked, ‘hidden’ for elements that are currently hidden from the user, etc.), and ‘properties’ of those areas (e.g., ‘autocomplete’ indicates when there’s a completion suggestion provided to the user, ‘required’ indicates that the element in a form is required for the user to complete before the form can be submitted). Because these tags are recognized by assistive technology, when used correctly, these tags can be very useful in helping users of assistive technology to navigate dynamically created web pages. However ARIA requires web page developers to be familiar with and willing to put in the extra effort necessary to add the tags to the web page when it is created and again if it is modified. When used incorrectly, the tags may actually hinder the user’s ability to successfully use the web page because the tag, recognized by the assistive technology, may cover the true purpose of the html element (for instance, by tagging a link element as a menu item).

5.1.1.3 Web Page Accessibility Checkers:

In addition to accessibility guidelines, the W3C also has an HTML markup validation service that checks a web page to make sure its HTML code follows the W3C guidelines so that it should work consistently on all major browsers (*https://www.validator.w3.org*). Without valid HTML and CSS code, there is no guarantee browsers will correctly handle the web page, nor that accessibility technology will work correctly with a web page. In addition, the validator flags code that doesn’t comply with basic accessibility standards (e.g., no ‘alt’ tags for images on the page) as invalid (an error).

While the W3C validator checks for valid HTML and CSS code, and does very basic checks for accessibility, a number of other validators have been developed solely for checking accessibility. The W3C has a list of over 70 accessibility evaluation tools for pdf files, email, mobile apps, and web page code. The web evaluation tools do everything from checking to see if a web site is compliant with the law in order to receive federal funding (http://www.508checker.com), to flagging HTML markups for accessibility and potential problems and suggesting areas where accessibility could be improved (http://accessibility-bookmarklets.org/ , https://cksource.com/ckeditor/services#accessibility-checker , http://wave.webaim.org/), to browser add-ons that give information about the accessibility of web pages based on the WCAG rules (https://addons.mozilla.org/en-US/firefox/addon/ainspector-sidebar/ , https://chrome.google.com/webstore/detail/axe/lhdoppojpmngadmnindnejefpokejbdd , https://chrome.google.com/webstore/detail/tenon-check/bmibjbhkgepmnehjfhjaalkikngikhgj), to developer's tools to help developers create accessible web sites (http://chromelens.xyz/), to tools that check for contrast levels and allow developers to see their web site as users with various forms of color-blindness would see it (http://gmazzocato.altervista.org/colorwheel/wheel.php , http://chromelens.xyz/, http://colororacle.org/, https://addons.mozilla.org/EN-US/firefox/addon/wcag-contrast-checker/ http://dasplankton.de/ContrastA/) , to tools for evaluating web pages for their cognitive level or grade level (http://www.online-utility.org/english/readability_test_and_improve.jsp , https://jellymetrics.com/readability-grader/) (*see https://www.w3.org/WAI/ER/tools/ for the complete list of W3C suggested accessibility evaluation tools*).

5.1.1.4 Social Accessibility Network

Even with all of the accessibility options available, many web pages still either do not have accessible code built in or have areas that are inaccessible or poorly accessible. In response to this, Takagi et al., (*2008*) proposed a Social Accessibility network, in which users who find web pages that are confusing and inaccessible can submit requests to a group of volunteers who, in response to requests, collaborate to develop suggestions for dealing with inaccessibility issues on web pages (*Takagi et al., 2008*).

5.1.2 Web Page Navigation

While web accessibility guidelines and checkers focus on assuring that the structure of the web page is accessible, which is clearly a basic prerequisite to successful acquisition of information from a web page, they do not help users to navigate content quickly and efficiently. As mentioned in the introduction section of this dissertation, a problem for many people using assistive reading technology is the rate at which they are able to process data using the technology. The problem of processing data quickly is more pronounced on many web pages because, in addition to singling out the main content (usually effortlessly identified by a visual reader), each web page might contain a section with links to other web pages or to content within the web page, author or company information, advertisement content, etc. Equally, the sheer amount of content on the web is overwhelming and difficult to process efficiently.

In order to process data more efficiently, many people who use screenreaders speed up the rate of playback. Screenreader users often report speeding up the rate of playback to up to 500 words per minute (as opposed to the average speaking rate,

which is about 180 words per minute). At rates of 500 words per minute, expert users are able to comprehend upwards of 50% of the content being read, although as the rate of speech goes up, the level of comprehension diminishes (*Stent et al., 2011*). Equally, even at speeds of 500 words per minute, the ability of a screenreader user to successfully process information about the content of a web page is still slower than that of a visual scanner (*Bigham et al., 2007, Takagi et al., 2007*). As a result, users of assistive technology have developed techniques to allow them to navigate more quickly through a web page.

Most browsers have built-in keyboard alternatives to allow users to navigate using keyboard shortcuts. Users often use the tab key to navigate quickly from link to link within a document, and the shift-tab key to go to previous links.

The screenreader Jaws from Freedom Scientific has built-in shortcuts that allow users to navigate specifically through a web page more efficiently. Jaws uses keystroke commands in order to navigate through the different HTML tags. Jaws allows the user to move forward and backward from header to header tag using the 'H' and shift-H. Users can navigate from paragraph to paragraph tags using the 'P' and shift-P key. Jaws allows users to choose to navigate to either the next already visited element, or, alternatively, the next unvisited element. Jaws also allows users to place markers within a document to be quickly navigated to in the future. In addition, Jaws has a mode called skim-reading, in which users can hear either the first line or the first sentence of each paragraph in a document. Unfortunately, according to Ahmed et al., (2012a) when interviewed, 20 experienced JAWS users either did not know about this feature or didn't use it because it was inconvenient and less than helpful.

None of the navigation options in Jaws are intelligently guided, although they do allow users to navigate through HTML elements that have been placed intelligently (most likely by the web page author) in a document *(JAWS 18 Documentation, 2017)*. These methods are not equivalent to the way in which sighted readers skim through information in a document.

In an attempt to create a more intelligent approach to scanning through a document, Ahmed et al., (2012 b) developed a method for summarizing a web page without skipping any content or losing the content order within the web page. The authors first determined what worked in terms of aiding nonvisual readers by having them develop a "gold standard summarization" when skimming through a document. Each subject summarized each sentence with only words in the sentence. The resulting summary contained words in the same order in which they occurred in the document, and the summary was at most one third of the length of the original document. The idea was to mimic ideal skimming through a document. They found that, contrary to most summarization techniques, in which the summary involved content that occurred most frequently within the document and skipped less relevant information in the document, the most useful approach to scanning was to reduce each sentence to the most salient nouns and word combinations within the sentence, preserving the overall order of the content of the document (Ahmed et al., 2012a). In this way all content and the order, or topology of the document was retained. To automate this summarization process, they evaluated each sentence within a web page to extract grammatical relations, then created a lexical tree, with each node being a word in the sentence. Grammatical and structural features of the sentence were used to train a classifier to determine whether words and word combinations should be used in the summary.

Thus each sentence was summarized individually for its content as represented by words and word pairs within the sentence. None of the content is skipped, and the content was presented in order of occurrence within the document to the listener. Users who wanted more information about a particular section of text within the document were able to switch between the scanning listening and the full-text listening through a keystroke. (Ahmed et al., 2012b). In Ahmed et al., (2013) the authors expanded their scanning techniques to allow users to create shorter and longer summarizations using a touch-screen interface.

In an innovative approach to allowing nonvisual readers to scan through information quickly, Guerreiro (*2016*) used separate audio channels (with different voices and at different speeds) to present to users different textual information being read aloud. They found that, especially when text was read at a slightly faster rate than the default rate (ideally around about 278 words per minute), users were able to focus in on relevant information from text faster than when listening to one audio channel at fast rates of speech. With two and three channels of speech, users were able to pick out a relevant sentence, in a manner similar to the "Cocktail effect" in which one is able to hear one's name at a cocktail party. While the approach taken currently doesn't work with screenreaders, users expressed a solid interest in adopting the technology because it allowed them to gather information about content faster than the alternatives they are currently using.

5.2 My Document Scanning Interface

For my interface I chose to use a web page written in HTML, CSS, and javaScript. A Web page has as an advantage that most people are familiar with web pages and how to interact with them. Web pages already have a great deal of

information on how to make them accessible (see Web Page Accessibility section), so basic guidelines for an accessible interface already existed. Equally, a Web page has as an advantage that it can be accessed anywhere in which a user has a computer and internet access. Once I had established that I would use a Web page as the basis for my interface, I needed to expand on the page so that it would allow access to the information users wanted quickly and efficiently, while continuing to work with accessibility technology.

My interface goals were similar to Ahmed et al., (2012a, 2012b, 2013) in that I did not want to lose the topology, or overall order of the document. Equally important, I wanted to allow users to switch back and forth between scanning mode and normal reading mode seamlessly. However, while I did not want to reword any of the document, my overall goal was to convey where in a document users might focus their attention when attempting to answer a question. My goals with my interface were also closely aligned with those of Guerreiro (2016) in that I wanted users to be able to quickly locate information relevant to a question or topic throughout the entire document. In developing the user interface, I wanted one that relayed the information to readers who used standard alternative technologies for reading. The user interface had to highlight text in the document relevant to the question.

Because of the way the MRI method worked, I had both what the method rated as the most relevant sentences as well as the most relevant paragraphs. Just conveying the sentences or paragraphs in the order in which they were rated, however, would've been an unfair disadvantage to users because they would have had no idea of where they were in the document for future reference (something visual readers clearly have), nor would they have had information about how much relevant information was

in a particular area within the document. Indeed, a big concern was allowing users to have control over how they navigated through the document so that they could gather information in a way that was most useful to them, not just in answering the question, but in gathering information about the question's topic and about the overall topology of the document. In addition, while each sentence individually often conveyed some useful information, the context in which the sentence appeared often included much more useful information. Equally, if many sentences that were ranked as highly relevant were located in close proximity to each other within the document, readers should be able to use that information to key in on that particular area within the document with the assumption that content in that area is most likely highly relevant to the question.

To this end I created a user interface designed to allow users to choose how they want to garner information. The user's initial introduction to the system is through the Access Page, which allows the user to type in a question and select a document, and then select a scanning mode (to be described below). Upon clicking a "submit" button on the web page, a server-side python script is called and the Document Page is generated, with the chosen text document modified to contain HTML tags necessary to make a web page, along with the appropriate html links (as described below) for navigating through the Document Page to access information related to the question. All web pages generated are validated using the W3C validator, and follow the WCAG 2.0 guidelines.

Both the Access Page and the Document web pages have four different buttons for the four different modes of navigation. By clicking on a button, a web page will be generated with the appropriate tags for different navigation modes (described below)

using a javaScript, and at any time users can click on any of the four different buttons to generate the Document web page in that navigation mode. I decided to generate a new web page rather than dynamically updating the existing web page because of the JAWS screenreader. JAWS traditionally stored a static copy of a web page, and until recently, did not automatically update the web page when content was dynamically updated. Because of this, many screenreader users turned off the automatic updates, citing as a reason their familiarity with static web pages (*Bigham et al.,2007*). To counteract this less than ideal approach to dealing with dynamically changing pages, I created scripts for the web page so that, when the user decides to change modes of navigation entirely, a brand new web page is created and loaded, rather than the content of the existing web page being dynamically modified. The web page includes the appropriate ARIA tags in order to make navigation easier for users.

I will first describe the initial user interface as given to users for feedback, and then I will describe the modifications made based on user feedback.

5.2.1 User Interface

The user interface consisted of an Access Page and a Document Page. The Access Page comprises a web page in which the user enters their question and the text document in which the question's answer should be found, along with mode in which the user initially wants to explore the document. The Document Page is a web page created from the original text document that contains all links relevant to the question and to the page's topology. Depending on the mode the user chose, different links will appear at the top of the Document Page. This interface, as described in detail below, was given to three consultants, two of whom were users of JAWS screenreader, and one who used screen magnification, for feedback.

5.2.1.1 The Access Page:

The Access Page consists of the following: an HTML text box, used to allow users to type in a question, and a browse box that allows the user to choose a document. Below this, the page has four buttons that allow the user to choose one of four methods for traversing the document. The four methods are: Sentence Mode, Paragraph Mode, Ordered Sentence Mode, and Topology Mode (and will be described in the Document Page description, below). In the original design, the user could roll the mouse over any of the four modes and a pop-up window would appear with a description of what the mode meant would show up (see Figure 3, below).

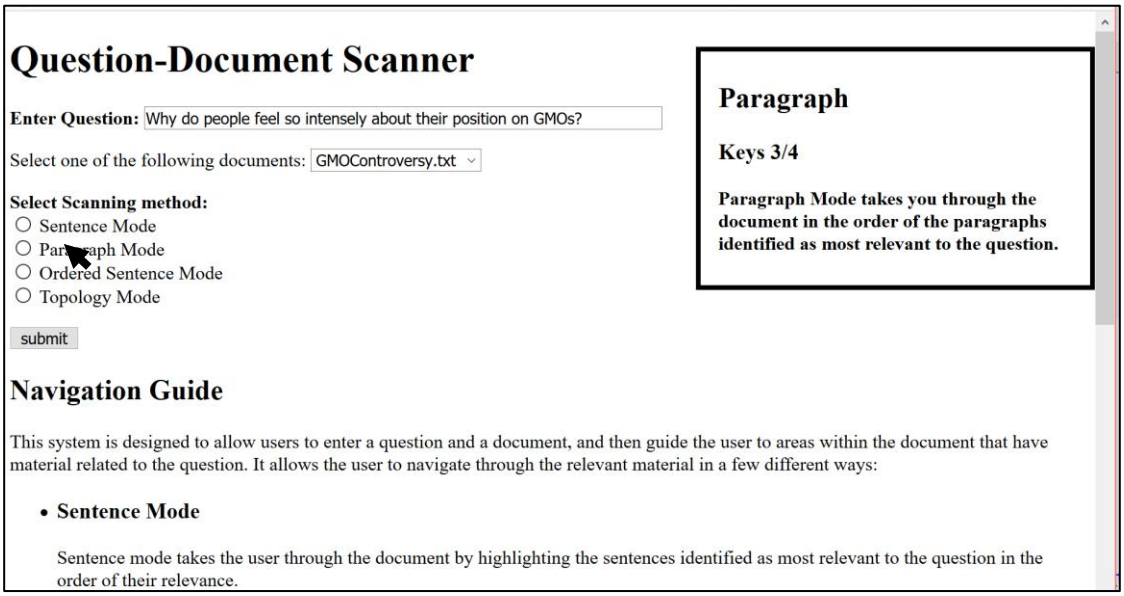


Figure 3 Screenshot of Original Access Interface with the user's mouse over Paragraph Mode and the Paragraph Mode Explanation pop-up box on the right.

Once a user entered a question in the question text box, the link to the document in the browse box (currently the document must be a text file), and chose a mode of navigation by clicking on one of the four mode buttons, the user then clicks on a submit button, which relays the question, document, and mode preference over to a python script on a web server that generates the Document Page.

5.2.1.2 The Document Page:

The Document Page consists of the original text document modified to be an HTML web page, along with all relevant links, anchors, and scripts based in part on the results of the user-entered question run through the MRI method and on the user's choice of scanning mode. The generated Document Page contains the question at the top of the page, followed by a set of links to locations within the document. The type of link is determined by the Mode choice the user made. This is followed by the original document text modified to contain relevant links and tags (ARIA, anchors, etc.) In addition, the content text in the Document page contains links between the linked sentences or paragraphs. Screenreader users and screen magnification users are able to jump from link to link within a document, so including these links within the content text itself allows users to scan through the document in the mode of their choice by jumping from link to link.

If at any time the user wants to switch modes, there are buttons along the top of the Document page for each mode. The user just needs to click on the mode of choice, and a new Document page is generated with links at the top of the page representing the new mode.

In addition, at any time users can switch modes within the Document page without generating a new page by using different key strokes. These keystrokes for

the different Modes work regardless of the Mode the user chose and thus the Mode represented as links at the top of the Document page.

The Mode choice was originally set up as HTML form radio buttons so the user could only choose one (as in the Access Page, see Figure 3, above). Once inside the generated page with the mode-appropriate links at the top, the user can switch between modes by either generating a new page in the new mode, or by simply switching to using different keystrokes that access the different modes (to be described in more detail, below). The Modes are described as follows:

- **Sentence Mode:** In this mode, a list of the sentences ranked as most relevant to the question will appear as links at the top of the Document Page. The ranked sentences (shown as the sentence, in its entirety) appear in the order of the ranking of relevancy obtained by the MRI method. Clicking on any of the sentence links at the top of the page takes the reader to the location within the document where that sentence occurred. In addition, the user can click on the ranked sentence within the document itself and be taken to the next ranked sentence within the document in the order of rankings (i.e., in the order of ranking, or the order in which they appear at the top of the document). Users can also navigate from ranked sentence to ranked sentence using keystrokes '1' and '2' ('1' will take the user to the next most highly ranked sentence, and '2' will take users to the previous most highly ranked sentence, based on where the user is in the Document Page. If the user has not clicked on any link and is not at a link within the text document, clicking '1' will take the user to the first ranked sentence, and '2' will take the user to the last ranked sentence based on the list order). While very little topological information is conveyed in this mode, users can still navigate to nearby relevant sentences, nearby relevant paragraphs, and to the beginning of the paragraph they are currently in simply by switching to using the different available keystrokes (being

described below) or by switching Modes entirely by clicking on one of the other 3 Mode buttons at the top of the Document Page.

- **Paragraph Mode:** In this mode, the paragraphs ranked as most relevant to the question by the MRI system appear as links at the top of the Document Page in the order in which they were ranked. The paragraphs appear as "Paragraph 25" "Paragraph 13", etc. Clicking on a paragraph link at the top of the page will take the user to that paragraph in the document. In addition, clicking anywhere within a ranked paragraph in the document will take the user to the next ranked paragraph in the document in order of ranking. Users can also navigate to next and previous ranked paragraphs using the keystrokes '3' and '4'. Again, while the topology is not preserved in this mode, users can at any time switch modes by switching keystrokes. Thus if users wish to find out what and/or how many sentences within and near the paragraph are ranked as relevant to the question, the user can switch to Ordered Sentence Mode keystrokes (described next). When Ordered Sentence Mode keystrokes are used while within a ranked paragraph, the user will be taken to the ranked sentence closest to the beginning of the ranked paragraph the user is currently in.
- **Ordered Sentence Mode:** In this mode, the sentences ranked as most relevant to the question using the MRI method appear as links at the top of the document as full sentences similar to Sentence Mode. However, in this mode, the sentence links are listed in the order in which they occur in the document. Thus if a ranked sentence occurs in paragraph 3, and another occurs in paragraph 4, and another occurs in paragraph 7, the Ordered Sentence Mode will take users first to the sentence in paragraph 3, then the sentence in paragraph 4, and then the one in paragraph 7, even if the sentence in paragraph 7 was ranked as more relevant than the sentences in paragraph 3 or 4. Users can click on any sentence in the list of linked sentences at the top of the document to go to that sentence within the document. Once the user has clicked on a link and gone to that sentence's location within the document, the user can then click on that sentence within the document

to go to the next ordered ranked sentence within the document. In addition, the user can use keystroke '5' and '6' to go to the next or previous ordered ranked sentence. At any time users can switch to either of the previous modes or the topology mode (to follow) by switching keystrokes. Using keystrokes '1' and '2' will take the user to the next ranked sentence based on rankings, using the current sentence's rank to determine the next and previously ranked sentence (so if the current sentence was ranked 4th most relevant, pressing '1' will take the user to the 5th most relevantly ranked sentence). Keystrokes 5 and 6 will take the user to the closest relevantly ranked paragraph. Thus, if the current sentence was in a ranked paragraph, pressing the key '3' would take the user to the first sentence in the current paragraph. Otherwise it would take the user to the next ranked paragraph.

- **Topology Mode:** In this mode, the ranked sentences and the first sentence of each paragraph appear as links at the top of the document in the order in which they occur in the document. As with the other methods, users are able to click on a linked sentence to go to the next linked sentence in the document. This method was designed to give users a general feel for the overall content of the document, as well as the overall topology of the document in relation to the question (so for instance, users could get a feel for when numerous ranked sentences occurred within close proximity). In addition to clicking on the links at the top of the web document and clicking on a linked sentence to navigate to topologically-ordered sentences, at any time, keystrokes '7' and '8' can be used to traverse the sentences in this mode.

At any time, in any mode, the user can switch modes entirely by pressing on one of the four Mode buttons at the top of the Document Page.

5.2.1.3 Key Strokes:

The links at the top of the Document Page and the links between linked elements in the content text will remain unchanged unless the user chooses to change Mode using one of the four Mode buttons on every Document Page, at which point a new Document Page is generated with a new set of top links and linked content elements. However, users can traverse the document at any time in any mode using the different keystrokes associated with the different Modes. As mentioned above, users can navigate in Sentence Mode using keys '1' and '2', Paragraph Mode using keys '3 and 4', Ordered Sentence Mode using keys '5' and '6', and Topological Mode using keys '7' and '8'. If users choose to switch modes, the keystroke will take the user to the closest sentence or paragraph based on where the user currently is.

In addition, users can at any time go to the first sentence of the paragraph they are in (for instance, when they clicked on a sentence link in Sentence Mode) by hitting the 'p' key. This allows users to go back to the beginning of any paragraph they think might be useful to read in more detail. At that point the user can switch to the keystrokes for Ordered Sentence Mode to determine the sentences in that paragraph that are related to the question. Alternatively, once at the beginning of the paragraph the user can switch to direct reading of the paragraph.

Users can also access a Help Guide at any time using the 'h' key. The Help Guide that explains each of the different modes, and which keystrokes can be used to access each of the modes.

In the original design, when the user ran their mouse over each of the Mode buttons, an explanation box popped up explaining that mode (see Figure 3, above).

5.2.1.4 Consultant Feedback

After the interface was designed and implemented, it was given to three consultants – two of whom used a screenreader (JAWS from Freedom Scientific), and one who used screen magnification (using the magnification system built into her computer). The feedback from the consultants who used screenreaders was notably different from the consultant who used screen magnification. This corresponds with Szpiro et al., (*2016*) who found that users with low vision had different issues accessing computers, tablets, and smartphones than users who were blind. Based on their feedback, certain interface set-up options were modified and added.

5.2.1.4.1 Screen Magnification Modifications:

The screen magnification user requested options that would make the interface easier to see visually, especially when the interface was magnified. Based on her feedback, the following options were added (See Figure 4, below).

1. A "Reverse Contrast" button was added to allow users to have white text on a black background. While reverse contrast is built into most browsers, it also reverses the colors on images. The added button only reverses contrast on text elements. This request by our consultant directly corresponds with problems encountered by low-vision subjects in Szpiro et al., (2016) who found that the reverse-contrast of images confusing and unhelpful.
2. Font Size '+' and '-' buttons were added to allow the user to increase or decrease the font size, as well as all web elements uniformly. While the screen magnification technology allowed the consultant to enlarge the text

as large as she wanted, it did not correspondingly increase the size of the elements in a form. For instance, radio buttons and checkbox buttons did not increase in size when the font size was increased, making it very difficult for the consultant who was low vision to select the desired buttons. The added '+' and '-' buttons allow users to increase both the font size and the form element button sizes simultaneously.

3. A Font Type option was also included, allowing the user to switch between 'Arial' and 'Times New Roman' font quickly and easily. The consultant preferred 'Times New Roman' but said that other low vision readers prefer a sans-serif font.

4. A Link Color option was added that allows users to switch the link colors to 'red', 'green', 'blue', or 'yellow'.

These settings were implemented so that the choices were maintained throughout the system's use. Once a user set these options they stayed set to the user's preference until the user changed them.

In addition to these options for allowing users to personalize the appearance of the interface, a change was made that excluded pop-up windows. In the original design, the user could roll his or her pointer over a particular mode button and a pop-up window would appear explaining the mode. In addition, the user could hit the 'h' key and a Help Guide Window would appear. However, because the user of text magnification magnified the text to such a great extent, the pop-up window covered a portion of the text within the document. Instead, a Help Guide Section was added to each Document Page and the Access Page at the bottom of the page. The Help Guide Section describes each of the modes and all of the keystrokes available. A "Help"

button was added at the top of the web page that brings the user to the bottom of each web page. Equally, the ‘h’ key takes users to the Help Guide Section. Finally, a button was added in the Help Guide Section that takes the user back to the main content of the web page.

The final change made based on feedback from the consultant using screen magnification was to change the Mode Buttons from a drop-down positioning to horizontal buttons. Again, because of screen magnification, the drop-down positioning took up too much space on the page. The consultant preferred the smaller footprint of a horizontal set of buttons.

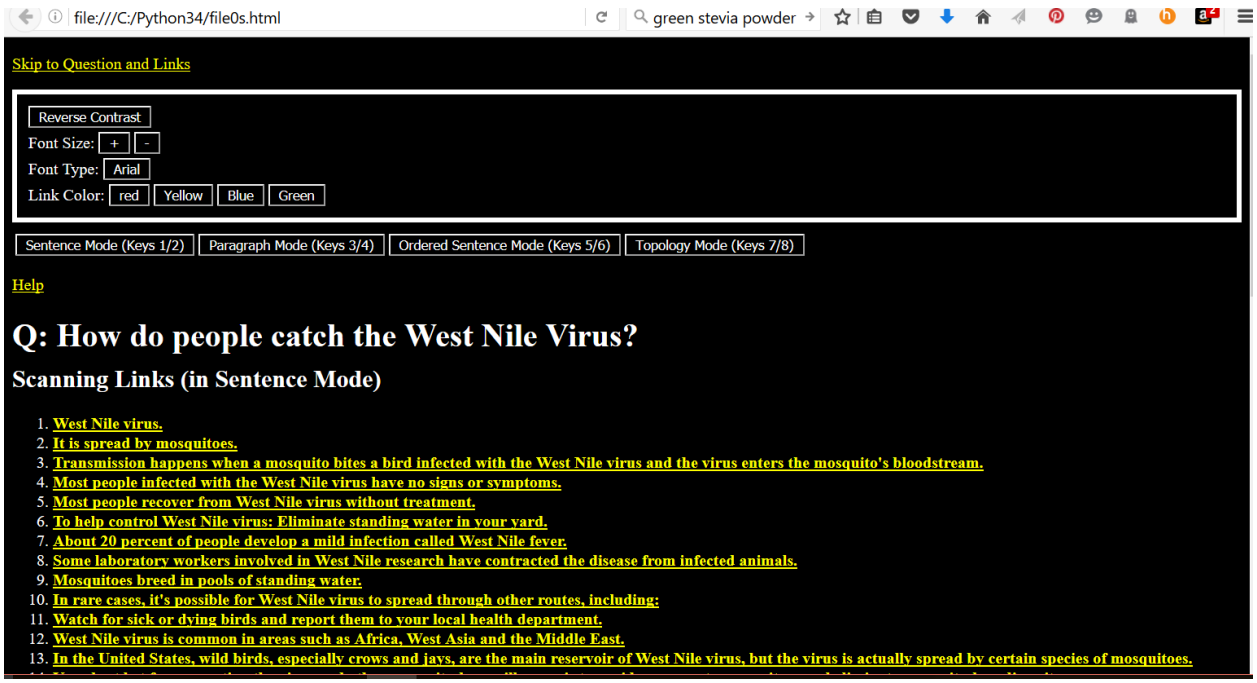


Figure 4 Screenshot of top of Document page in Sentence Mode, with Formatting buttons and alternative Mode buttons

5.2.1.4.2 Screenreader Modifications:

Feedback from the consultants who used a Screenreader indicated that some of the keystrokes didn't work in conjunction with JAWS. Consulting with Freedom Scientific (the manufacturers of JAWS) revealed that currently all keystrokes are used for different purposes by JAWS. However, if a software system wishes to override the predefined keystroke meanings for certain keys, the software can include code that indicates to JAWS which keys to reserve for a particular system (this override is employed by Facebook and Twitter, among many other software programs). Thus the system was modified to include the override code, reserving keys '1', '2', '3', '4', '5', '6', '7', '8', 'h', and 'p' for my system. This allows these keystrokes to work with my system for users of JAWS with Google Chrome, Microsoft IE, and Mozilla Firefox. However, according to the company (and confirmed by one of the JAWS consultants), JAWS does not support Apple's Safari and thus, while the system in general will work on Safari, the specific keystrokes will not.

After the modifications requested by the users of screen magnification were added to the user interface, the consultants who use a screenreader requested a "skip navigation" link that would allow the screenreader users to skip over the section on the web page that allowed users to set the page up visually. This link was added as well.

5.3 Evaluation of the System

Once the user interface was developed and vetted by consultants to their satisfaction, an experiment was set up to evaluate the efficacy and benefits of using the system. The goal of the experiment was to evaluate whether the system both worked well in the manner in which it conveyed information and also whether the information

being conveyed was useful in gathering information about a question that would aid in answering the question.

For the experiment, 10 documents with sentences and paragraphs generated using the MRI system were compared to the same 10 documents with the same number of sentences and paragraphs, only with these sentences and paragraphs chosen randomly throughout the document. So, for instance, if a document had 13 ranked sentences and 8 ranked paragraphs using the MRI system, the random version of the same document would have 13 randomly chosen sentences and 8 randomly chosen paragraphs. While both versions allowed the user to use the interface, and thus granted the user a certain amount of topological information about the document, I theorized that comparing randomly chosen paragraphs and sentences versus sentences and paragraphs chosen intelligently using the MRI system would give feedback on the usefulness of intelligently identifying relevant information within the document, and, specifically, the MRI system's usefulness in intelligently identifying relevant information and its synchronization with the user interface.

5.3.1.1 Hypotheses:

1. Subjects will find the MRI system quicker and more efficient to use when seeking relevant information in order to answer a question than they will using a system with randomly generated information.
2. Subjects will find the MRI system more helpful in locating relevant information than they will using the system with the randomly generated information

5.3.1.2 Subjects

Subjects participating in the study had to be users of some form of assistive technology for reading, 18 years of age or older, and native speakers of American English. Subjects were recruited through contacts who used assistive technology for reading. The three consultants were asked to contact friends and acquaintances who used assistive technology for reading to participate in the study. In addition, the study was posted to the DHSS mailing list, the BlindPhiliComp technology group, and specific blind technology groups that the consultants had access to through their work. Subjects were also recruited through a local optometrist who conducted research, and through online neighborhood groups. The message posted was as follows:

> Please consider helping in the design of a system for users of assistive reading technology such as a screen reader (e.g., JAWS) or a screen magnifier if you are a regular user of those technologies. The goal is to create a system that will help people who use assistive reading technology to answer homework and work-related questions more quickly and thoroughly. For your participation you will receive a $10 Amazon gift card.
>
> The study will compare some different systems that locate areas of text that you might want to read in order to answer a given question – the methods used by these systems are different and we anticipate that some might be better than others. After having a chance to play with the system, you will be given one of the systems, and be asked to use it to answer questions about relatively short (at most 2 page) documents. After you have completed this, you will be asked a few questions about the usefulness of the system. At the end of the study you will be taken to a page where you can enter your information to receive the $10 Amazon gift card (your identifying information will in no way be associated with your study responses).
>
> The study can be done at home on your own computer. It should take about an hour, although you can stop and come back to it any time over the course of 2 weeks.
>
> The study can be located at: http://dyphd.agora-net.com/v2/ScanningApproval.html

5.3.1.3 Task Description

The study involved having subjects use the system online (described below in more detail) with 10 questions and documents. Subjects were required to scan through 10 short documents (equivalent to two pages in length) to determine the answer to a question related to the document. Once subjects scanned through the 10 documents and answered the corresponding 10 questions, they were asked a series of questions about the value and ease of use of the system. Subjects could use any form of assistive technology they were comfortable using, and, because the study was online, subjects could use the type of computer and browser they were most comfortable with (although JAWS users were advised not to use Safari because Safari is not supported by JAWS).

Upon going to the online study link, each subject first read a consent form. Upon agreeing by clicking on the "I Agree" button, users were sent to a system description page that described the experiment and the different modes. From this page users were able to link to a training web page which was a two page document in the form of a web page with the links and options that make up the user interface. Users were able to spend as much time as they needed exploring the training document web page until they felt familiar with and comfortable with the web page interface and the different modes for traversing the document web page. In this web page (as well as in the consent form page and the system description web page) users had the options for modifying the visual appearance of the web page and subsequent web pages as described in Section 5.2.1.4.1. Subjects were allowed to train on the system using the training document web page for as long as they wanted to, including switching modes as often as they wished.

Once subjects were comfortable with the system, they were then given 10 document web pages sequentially, each with a relevant question. Each document web page was in the form of the User Interface described in Section 5.2.1. The only difference was that a prechosen question appeared at the top of each document web page, with four potential multiple-choice answers beneath the question. Thus, rather than have subjects enter a question and a document, and have the document web page generated (which has the original question but no multiple choice answers appear at the top), subjects were given the document and the question in order to ensure the comparison of equally challenging questions and document topics. The multiple choice answers were included to make sure subjects were motivated to find an answer, and, equally, to make analysis of results straightforward. Subjects were instructed to scan through each document web page in any mode they chose, switching between modes if they chose, in order to answer the question. Once a multiple-choice answer was chosen, the next document web page with pre-selected question appeared. Document web pages were presented one at a time in random order, and subjects were only presented with the next document web page when they had answered the question associated with the previous document.

Each subject was assigned randomly to either the MRI group, in which the document web pages used the MRI method to determine the most relevant sentences and paragraphs, or to the random group, in which the document web pages generated had randomly chosen sentences and paragraphs. For those in the random group, for each subject a new set of random sentences and paragraphs was generated to ensure that no quirk resulted on a set of highly-relevant sentences and paragraphs being generated and used continuously by all the random subjects. Exactly the same number

of sentences and paragraphs were chosen with both the MRI method and the random method.

At any time throughout the experiment, subjects could adjust the appearance of the web page document by changing contrast, font size, link color, and font type. Once adjusted, the appearance would stay that way until subjects adjusted any of the appearance features at a subsequent time.

When a subject had finished going through the 10 document web pages and had answered the question associated with each web page, the subject was then taken to a Qualtrics survey in which the subject answered questions about the efficiency and ease of use of the system. The survey included questions like, "Did you find the system more or less helpful than not using the system in answering the question?" and "Do you think using the system allowed you to answer questions more or less quickly than not using the system?" (see Appendix C for the entire survey)

It was anticipated that most subjects would be able to complete the entire experiment (i.e., answer all 10 document questions and the meta-questions about the system) within an hour. However, in order to accommodate subjects who fatigued easily or took longer than anticipated, subjects were allowed to take up to two weeks to complete the experiment. They could leave the experiment at any time and return at a later date. When they returned the study would pick up where they left off.

Once subjects had completed the survey, they received a $10 Amazon gift card for their participation.

5.3.1.4 Results

Seven subjects completed the study. Four of the subjects used the system with MRI-generated sentences and paragraphs, and three subjects used the system with

randomly-generated paragraphs and sentences. Five of the subjects used a screenreader, and two of the subjects used screen magnification.

None of the results were significant.

Results of the questionnaire can be found in Appendix D

5.3.1.5 Analysis of Data

Most questions in the survey can be loosely divided into three categories: those comparing the efficiency of the system with MRI-generated data versus the system with randomly-generated data; those reflecting the general helpfulness of the system; and those dealing with the system's functionality. Questions 1 and 2 reflected the efficiency of the system, Questions 4, 6, and 7 reflected the general helpfulness of the system, and Questions 3a,b,c, 5a and b, 18, and 21[6] dealt with the system's functionality.

For the questions that reflected the comparison of the MRI-generated data versus the randomly generated data, there was a slight positive inclinations toward the MRI system. For instance, for Question 1, "Did you find the system more or less helpful than not using the system?" subjects were given the choice of, "a lot less helpful", "somewhat less helpful", "neither less or more helpful", "somewhat more helpful", or "a lot more helpful". All of the subjects using the system with MRI data rated the system as being either "a lot more helpful" or "somewhat more helpful". For those using the system with random data, none rated the system as "a lot more

---

[6] There were 12 questions in the Survey. The numbering was done automatically by the Qualtrics Survey System. After trying to change the numbering, I decided question numbers were not relevant to the information gathered by the survey, and left them as the Qualtrics System numbered them.

helpful", and one rated the system as "somewhat less helpful". If one rates the answers on a scale of 1 to 5, with 1 being "a lot less helpful", and 5 being "a lot more helpful", the average for the MRI users was 4.25, whereas the average for the Random data was 3.67 (see Figure 5, below).

For Question 2, "Do you think using the system allowed you to answer question more or less quickly than not using the system?" the trend was equally favorable towards the MRI system. With possible answers ranging from "a lot less helpful" to "a lot more helpful", again, the answers for the MRI system were either " a lot more helpful" or "somewhat more helpful", whereas answers from subjects using the randomly generated data ranged from "somewhat less helpful" to "somewhat more helpful" (see Figure 5, below).

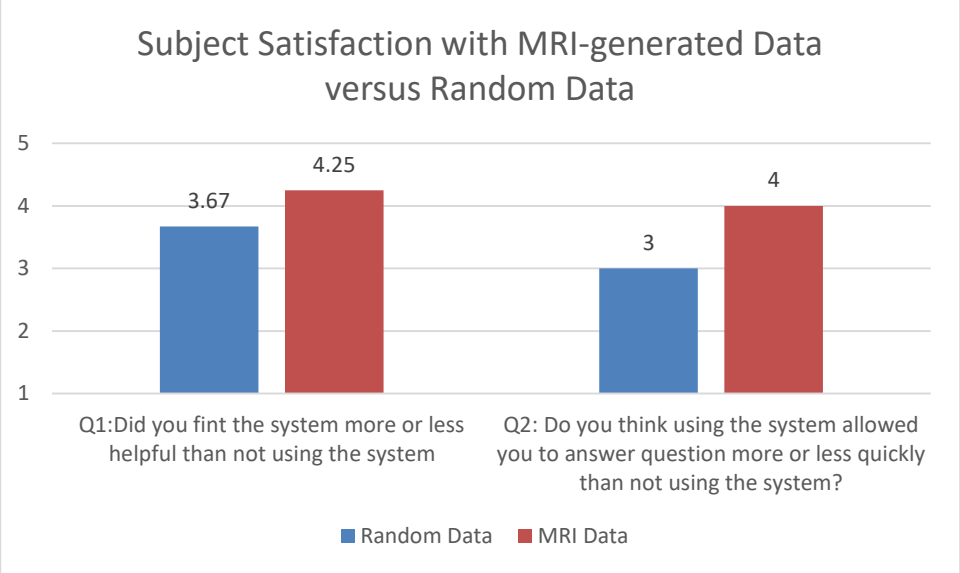


Figure 5 Comparison of results of questionnaire on use of system with MRI data versus randomly-generated data.
5 = "a lot more helpful","a lot more quickly" respectively.
1 = "a lot less helpful", " a lot less quickly" respectively

For the questions related to the general helpfulness of the system, the results also trended towards positive. For Question 4, "Did you find the system confusing or straightforward to use?" subjects who received random data rated the system as somewhere between "somewhere in the middle" to "somewhat confusing" with an average rating of 2.67. In the questions related to the system's helpfulness, this was the only averaged answer below 3. Subjects who received the MRI data, on the other hand, rated the system as a 3.75 (between "somewhere in the middle" and "somewhat straightforward"). (See Figure 6, below).

For Question 6 and Question 7, "Would you use this system again to help find information related to questions in documents?", and "Would you recommend this system to a friend?", respectively, the subject responses again averaged positive. The average answer for Question 6 was between "Definitely yes" and "Probably Yes", with subjects who received random data more likely to choose "Definitely Yes" (with an average score of 4.67 for the subjects with random data and 4 for subjects who received the MRI data – see Figure 6, below). For Question 7, subjects who received the random data again chose either "Definitely yes" or "Probably yes" with an average score of 4.33, and subjects who received the MRI data had an average score of 3.5, with the lower score reflective of one user who answered "Definitely Not", which was quite an anomalous answer based on this subjects previous answers and thus possibly a mistake.[7]

---

[7] This answer was anomalous to this subject's answers to previous questions, which included answering "Definitely" to "Would you use this system again?" "Very Straightforward" to "Did you find this system to be confusing or straightforward to use?", "A lot more" to "Did you find this system more or less helpful than not using the system", and "A lot more" to "Do you think using the system allowed you to answer questions more or less quickly than not using the system?" Thus answering "Definitely Not" to "Would you recommend this system to a friend?" was glaringly anomalous.

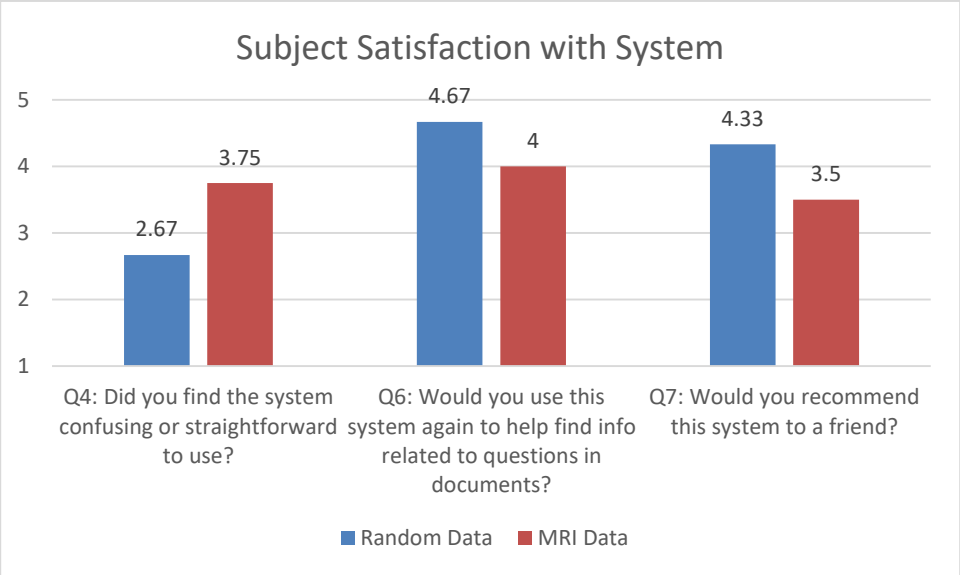


Figure 6 Subject satisfaction with system measured by subject response to questionnaire.
1=”Very confusing”,”Definitely Not”, “Definitely Not” respectively
5=”Very straightforward”,”Definitely,”Definitely” respectively

For the questions related to the system’s general functionality, there were no technical issues, and only a few complaints or suggestions. Equally, subjects seemed to be able to use the system without a great deal of training. When answering Question 3a, “Did you use more than one scanning technique in answering the questions?”, 5 of the 7 subjects reported using more than 1 mode (and one of the subjects who answered “no” then reported using both the topological mode and the sentence mode in the subsequent question about which modes the subject used). Figure 7, below, shows the Modes reported to be used by the subjects in answer to Question 3b, “If you used more than one, which modes did you use?”. Sentence Mode, the Mode in which the most relevant sentences are listed and linked to each other by

the order of their ranked relevance to the question, was the mode used by the most users, followed by paragraph mode, then ordered sentence mode, and the least popular mode was the Topological Mode. Since Sentence Mode was the default mode the system was in when users started the experiment, it is logical that that mode would be the mode used by all 6 of the subjects who reported using more than one mode. Four of the six subjects who reported using more than one mode reported using the Paragraph mode, which allowed users to traverse among the most relevant paragraphs in the document, in order of most relevant to least relevant. Three of the subjects reported using the Ordered Sentence Mode which listed and linked the most relevant sentences by the order in which they occurred in the document, as opposed to ordering them by their relevance ranking, while two of the subjects reported using the Topological Mode, which was identical to the Ordered Sentence Mode with the exception that the first sentence of every paragraph was included in the list and links of sentences.

The only slightly notable difference in the choice of modes used based on whether the subject received randomly chosen data or MRI-chosen data was their use of the Paragraph Mode. Three of the subjects with the MRI-chosen data used the Paragraph Mode, whereas only one of the subject with the randomly-chosen data used the Paragraph Mode. For both groups, one subject chose the Topological Mode, and two subjects in the MRI group used the Ordered Sentence Mode and one subject in the random group used the Ordered Sentence Mode.

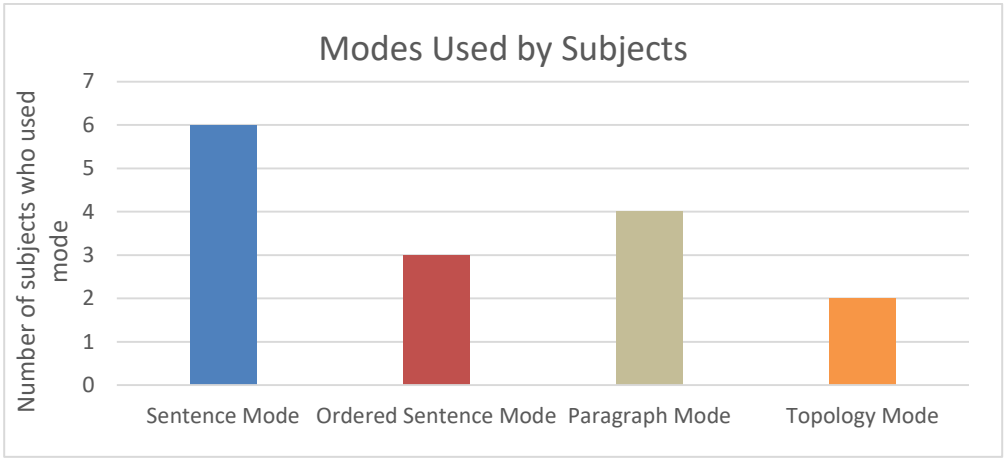


Figure 7 Graph of the number of subjects who reported using the different modes available in the system.

For Question 3c, "If you used more than one mode, which mode(s) were most helpful?" the mode reported as most useful by 5 of the 6 subjects who answered this question was the Sentence Mode. One of the 6 subjects chose Paragraph Mode as well (subjects were able to choose more than one mode for this question), and one subject reported that "all were equally helpful". See Figure 8, below, for a chart of the mode(s) chosen as most helpful.

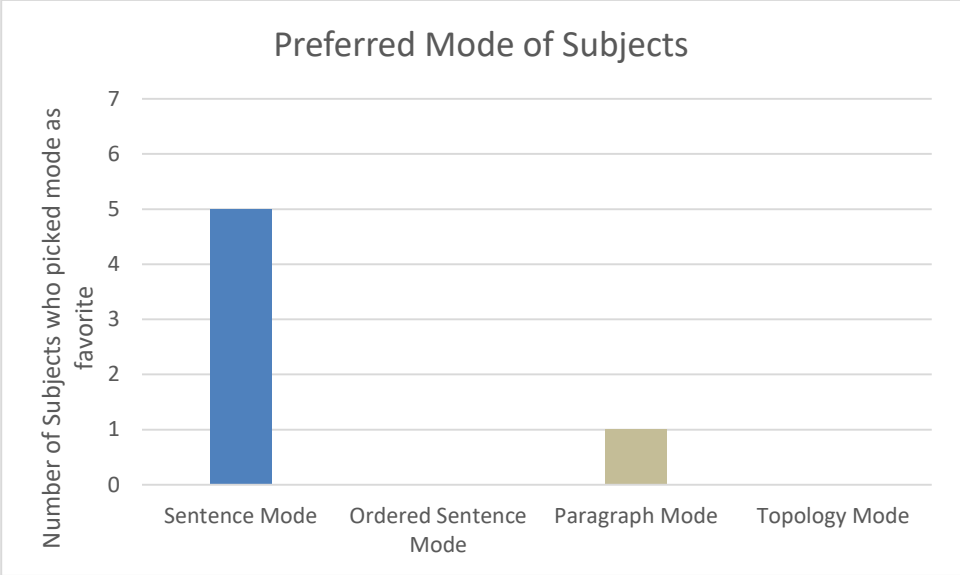


Figure 8 Graph of the number of subjects who reported a mode as most useful

Few problems and suggestions for improvements were included in subject responses. In answer to Question 5, "Did you experience any technical difficulties?" two subjects reported technical difficulties, but in Question 5b, "If you experienced technical difficulties, could you please describe them?" one of the subjects who reported technical difficulties answered "None" to this question. The other subject who experienced technical difficulties responded, "I use Jaws 18 and it was unclear at times if the hot keys in the system were actually changing the behavior. I think this was more a function of Jaws than of the system though."

Along the lines of Questions 5a and 5b, Question 18 asked, "Can you suggest keystrokes to replace existing keystrokes used for moving around the web page?" three subjects responded with "no", "none", or left the question blank. One subject said, "the numbering was easy, stick with it." One subject suggested "Perhaps use single letters to correspond to the modes; S, O, P and T", which were the original keys

chosen for the different modes, but, based on feedback from the consultants that those keys were frequently used by the JAWS screenreader , were switched to the numbers. One subject suggested, “arrow keys they are closer to the edge of the keyboard”, which also make sense but in most browsers (Firefox, Chrome, Safari, etc.) makes the web page bounce up or down, and thus I felt it would be quite disturbing to users of Screen Magnification technology. One subject said, “I am not sure but the numbers conflict with standard keys, so if people do not go into the settings to change this as described, or it gets messed up, it won't work. Perhaps have it verbalize something when switching between modes.” While I was aware of the conflict, and had had a discussion with Freedom Scientific about the fact that screenreader users would have to change the settings, the idea of presenting an aural message before users begin using the system (at least the first time) seems like a very good suggestion and will be taken into consideration for future updates.

In answer to Question 21, “Do you have suggestions for options that would enhance the usability and usefulness of this system?”, one subject responded, “Looks good, keep it up!” and another responded, “Have it show the current mode so it could be arrowed over.” This seems like an excellent idea that is easy to implement and will be taken into consideration for future updates. The final question, “Any additional comments?” garnered two notable responses. One subject said, “While this comment is not related to the system itself, I did find that I tended to default to screen reader,” and another said, “Reading is always a challenge as well as dissecting the information. Once I got used to it, it was easier to locate the relative information and move on. A true time saver and with my eye strained induced headaches, a lot less painful. Thank you.”

5.4 Discussion

Clearly, the number of subjects who completed the study was disappointing. The sparsity of responses made it notably difficult to get any significant data results. Equally, results that may have been in error were not ameliorated by large quantities of data. The question becomes, why were there so few subjects who completed the study?

Upon examination of generated data, it became evident that at least 22 people who received the random data started the study but failed to complete the study. I was able to determine this because for the randomly generated data, every time a new subject who received random data started the study, a new set of random data was created. Since the study was released, 25 sets of random data were generated. Three subjects who received randomly-generated data completed the study. Thus 22 people started but did not complete the study. Unfortunately since the MRI-generated data was the same for all subjects who received the MRI-generated data, there is no way of telling how many subjects who received the MRI-generated data started the experiment but failed to finish it. This is unfortunate, because comparing the number of subjects who finished who had the MRI-generated data versus the randomly-generated data might have been interesting in terms of the frustration of use of the system with MRI-generated data versus randomly-generated data.

In order to get feedback on why so few subjects completed the study, I enlisted the aid of a low-vision consultant I had not used before[8] to complete the study and give feedback on its ease of use. His feedback was that the study was extremely tedious and boring. Based on this feedback and the number of incomplete studies, it is

---

[8] My husband.

apparent that the $10 Amazon card was not enough of a motivator for subjects to complete the study. In the future, I would both decrease the length of the study and increase the reward given for subjects who completed the study. In addition, I think it makes sense to either go to subjects or have subjects come to my location to complete the study. I think subjects would be more likely to complete the study if we had scheduled a time and place for them to complete it. Equally, I may have been able to garner information about how subjects were using the system if I had been present as they were using the system.

Even with the sparsity of data, there were some encouraging trends. For both the questions related to whether the system was helpful and whether the system was useful in answering questions, subjects who received the MRI-generated data rated the system more positively than subjects who received the randomly generated data. While the data was not significant, the fact that there was any positive trend at all was encouraging, especially considering that simply by using the system, all subjects received links throughout the system that gave them a general overview of the content of the document and thus could lead to responses that the system helped them with finding content related to the question. Thus the fact that there was any difference at all and that it was consistently in favor of the MRI-generated data leads me to believe that the MRI-generated data was helpful.

Another positive trend, while again not significant, was the fact that most subjects used more than one mode to navigate around the system. This is encouraging because it indicates that subjects were aware of and comfortable with switching between the different modes, even after very little training and experience with the system. Most subjects chose the Sentence Mode as their preferred mode, which, while

it is the default mode, is also the mode that brings subjects directly to sentences ranked as the most relevant to the question, and thus this mode makes the most sense in a short document. While it does seem to indicate that subjects understood how the different modes worked in their choice of this mode, it is possible that this was the subjects' preferred mode because it was the default mode the system started in, and thus the mode subjects were most familiar with. Interestingly, the second-most used mode the Paragraph Mode (a mode used four of the subjects and chosen by one subject as the most useful mode), which took them to paragraphs ranked as most related to the question. Both of these modes dealt with data that was ranked as most related to the question, as opposed to the other two modes which both included data ordered by its topological order. Subjects seemed to prefer going directly to data ranked most relevant to the question over traversing relevant data in a way that gave them a feel for the overall topology of the document. This might be because users of screenreaders have other methods available to them to garner information about the general topology of the document. Again, what is encouraging is that subjects seemed to grasp the difference between the modes and purposely chose modes that took them directly to data ranked most relevant.

Finally, it was a relief that there were very few reported technical difficulties experienced by the subjects using assistive reading technologies. While most subjects reported being users of the JAWS screenreader, one reader reported using NVDA (a screenreader included with Microsoft Windows) and one subject reported using Voice Over (a screenreader included with Apple's Macs). In terms of screen magnification, one subject reported using ZoomText, (a screen magnifier included with Microsoft Windows), and one subject reported using Apple's Screen Magnification. Thus there

was a wide range of technologies being used by the subjects. That there were no real technological difficulties speaks well of the technologies, as well as their ability to interface with web pages.

The suggestions for alternate keystrokes for switching between the different modes were not particularly helpful. That said, it was known while designing the system that every keystroke is now used for some purpose by the Jaws Screenreader, and thus I would have to override some existing keystroke associations. The conclusion I came to based on the feedback received from these subjects is that the keystrokes picked for this system (the numbered keys) were not interfering with fundamental keystrokes used at high volume by any of the users, and thus probably as good if not better than other keystrokes I could have picked.

Finally, the suggestions made were good ones and relatively easy to implement. The suggestion for including aural instructions for turning off the keystroke override built into JAWS was a good one. In consulting with Freedom Scientific, I was assured that, because every browser overrides JAWS built-in keystroke associations for certain keys, almost all users of JAWS have not set up their screenreader to not allow the overrides. However, clearly one of my subjects had set up JAWS to not allow the overrides. Including aural instructions, or at least a warning that JAWS be set up to allow overrides in order for my system to work, seems logical and relatively easy to include. Equally, one subject suggested a button or box on each page that subjects could arrow over that would tell users the mode the user is currently in. Again, this seems like an excellent suggestion that can easily be implemented.

Overall, while there was a disappointingly sparse amount of data generated from the study, the results that were generated were to some extent positive and

encouraging. Equally, the study served as an interesting and informative lesson in how not to conduct a study and what needs to be taken into consideration for future studies.

Chapter 6

FUTURE WORK

This dissertation can largely be broken down into three components: ascertaining whether visual readers make semantic connections when scanning through documents for answers to complex questions; identifying a method for replicating the connections that visual users made during the scanning process; and creating a user interface that worked with assistive technology to allow users to scan through documents in an efficient manner when answering questions. While the overarching goal of the dissertation was to demonstrate the feasibility of replicating the process of scanning through documents for text visual readers identified as related to a question, each of the individual components lends itself to potential future work.

While the data collected from the scanning experiments is interesting, there is a good deal more to be learned from the results obtained from the scanning experiments. At the time of the eye gaze experiments, the Tobii Eye Tracker System had relatively limited fine grain analysis. Thus determining exactly where a person was gazing for longer periods of time could be somewhat off, by as much as a line of text. Finer grade analysis should reveal more specifically what caught the eye of the scanners. While this would require repeating the original scanning experiments in which visual scanners scanned through documents looking for question answers, the improvements in the Tobii Eye Tracker System might indicate exactly what caused the user to stop the quick "zigzag" scanning process to focus longer on a particular area of text. In narrowing down exactly what word or words caused a user to focus, the process of creating the semantic connections could be refined as well, allowing for more accurate analysis of successful replication of connections.

An important area of future work would be to repeat the eye gaze experiment with longer documents. While the scanning results did show that users do make semantic connections when scanning through a document for answers to a question, it is notable that in all likelihood this system would be used most frequently with longer documents. The benefits of this system would hopefully be notably more noticeably in a 50 page chapter than in a two page paper. Thus it is important to see whether users scanning through longer documents use the same or similar techniques for identifying relevant text as they do for short documents.

The work in this dissertation also showed that it is possible to automatically identify areas of text visual users identified as relevant to a question. A key component of this process was collecting word clusters by implementing Google searches on keywords in questions, then using the resulting links and their accompanying description snippets to link to relevant web pages and collecting the words surrounding the snippets in the web page. Shortly after the research for that portion of this dissertation was completed, Google discontinued its authorization for individuals to access and use their search API. At this point, no acceptable alternative has been found that allows for the gathering of relevant web links and garnering related word clusters. More research is needed to find a free, unlimited search system that allows for the creation of semantically related word clusters.

While the MRI method for identifying semantically relevant text was quite promising it currently does not take into account physical cues, such as first paragraphs, first paragraphs on pages, titles, lists, bolded words, etc. If the system really has as its goal giving the user all information visual readers glean when

scanning with the same level of priority, it should incorporate these physical cues into its prioritizing of sentences and paragraphs.

In addition, it would be interesting to continue to explore methods for automatically partitioning segment boundaries by looking at where the local topic shifts within a document, although a preliminary pass with Text Tiling (Hearst, 1993) resulted in segments too large for a refined result search in the single-topic documents used for this study. The analyses in this dissertation suggested that users focus on paragraphs in which there has been a topic shift, especially when the shift is towards a topic more closely aligned with the topic of the question whose answer is being scanned for (e.g., when the topic switched from discussions of statistical uses of Marijuana to the chemical make-up of Marijuana when answering the question, "How does Marijuana affect the brain?"). Because the MRI system has as one of its goals to give the user an overall feel for the content of the document, as well as information similar to that which a visual reader gets when scanning through a document, it would be beneficial to be able to incorporate topic shift markers within the system, possibly incorporating that information into the different Scanning Modes or a new Scanning Mode.

As mentioned in section 4.4.4, since the research on the MRI method for identifying semantic relatedness was conducted, was the state of the art has advanced and it may be beneficial to research the benefits of newer techniques such as word2vec and GloVe. It might be that incorporating these methods of identifying the semantic relatedness of words could very well improve the process of automatically identifying text visual scanners were likely to focus on.

Finally, the user interface was developed using user-centered design and worked well for the consultants and the experiment subjects. It would be valuable to continue working with consultants to improve the user interface. Currently the system is a basic web page. Because it is largely text based, the page scales relatively well to a smartphone. However, no real work has been put into optimizing the system for a smart phone. Clearly in today's world the system almost has to be smartphone-accessible. By optimizing the system to allow for flexibility between a computer and a smartphone, the system will be more mobile and thus useful for today's users.

In addition, it would be interesting to explore the use of automatically generated audio markers throughout the document that would allow users to use particular keystrokes to get information about, for instance, what paragraph they're on (out of the total number of paragraphs) in the document. Currently the system does not take advantage of the potential of audio channels in giving the user information.

The current interface was designed based on feedback from users of reading technology who were blind or low vision. One important future direction would be to work with users of assistive reading technology who have a learning disability such as dyslexia. Of the estimated 20% of individuals in the United States with a learning disability[9], 85% have their problems in the area of reading, including dyslexia (National Institute of Child Health and Human Development, 2007). While I am assuming that such a system could be equally invaluable to these individuals, it is

---

[9] According to the National Institute for Literacy (2007), a learning disability is "A severe difficulty in learning to read, write, or compute. Those with a learning disability have a significant discrepancy between what is expected of them given their general level of cognitive ability and their actual reading, writing, or mathematical ability or achievement. They may also have significant listening or speaking difficulties. Their difficulty is not due to mental retardation, social or emotional problems, sensory impairment (such as severe vision problems), or environmental factors (such as poor schooling)."

possible these users may use the system in a way we hadn't anticipated and thus it is critical to work with people with dyslexia and other reading disabilities to create a user interface specific to their particular needs.

Finally, the experiment conducted to evaluate the system and, in particular, the user interface, was clearly disappointingly small and it would be valuable to run the experiment again taking into account different lessons learned. For example, the experiment was relatively long and not terribly interesting. It would be beneficial to shorten the duration of the experiment before asking for feedback. It would also be interesting to move the experiment from an on-line experiment to an in-person experiment to get immediate feedback and to be able to glean information from watching users interact with the system. Since the expectation is that the system will be most beneficial with longer documents, it might be interesting to run a timed experiment in which half the subjects had to answer a particular question in a longer document using only their assistive technology and the other half would answer the question using both their assistive technology and the system.

Chapter 7

CONCLUSION

The goal of this dissertation was to create a prototype system for allowing users of assistive reading technology to scan through documents to efficiently locate answers to complex questions. In order to accomplish this, I had to successfully show that visual users scanning through documents did use as cues semantic relatedness when choosing areas of text o focus on in a document, I had to show that it was possible to automatically replicate the semantic connections these visual scanners were making, and I had to create a user interface that was user-friendly for people who were blind and for people who were low-vision and that user interface had to successfully convey the information visual scanners were getting when scanning through a document to these assistive-technology users. Each of these goals was accomplished.

The work completed for this dissertation has several contributions, the largest of which is a prototype of assistive technology that enables a person who is blind or visually impaired and uses assistive reading technology to scan a document, gaining information a visual scanner can get in a similar amount of time. The system will allow blind and low vision users of assistive reading technology to answer homework questions in a manner that is more efficient than they are able to with the tools they have available to them currently. However, the system will also allow users to locate relevant text more efficiently and even re-locate text they already read through and want to go back to more efficiently than techniques available with existing technology. To my knowledge there is no technology available today that offers similar benefits and it could make a significant difference in educational and employment

achievements by this population. Equally, it is hoped that the system could easily be expanded to work for people who use assistive technology to read because of a learning disability, such as dyslexia. Thus this system has the potential to significantly affect the quality of life for the large population of users who use assistive reading technology.

A significant portion of the work involved NLP techniques for identifying relevant information. While this portion was completed over five years ago, and the field has made significant developments since then, at the time of completion this work, inspired by Question-Answering and Query-biased summarization work at the time, made significant strides beyond these techniques. It is hypothesized that the MRI technique used here could be used in conjunction with current techniques to improve the discovery of semantic relatedness for areas such as Query-biased summarization since it is identifying areas of a document most relevant to a query in a method that goes well beyond the simple matching in current use and does not require any external knowledge base or pre-computing based on particular topics or queries. In addition, the MRI method is unique in that it attempts to replicate the semantic connections people seem to make intuitively when answering a question. Thus the information returned may or may not be the information most relevant to the question, but is more likely to be the same type of information that people would get from a document when answering a question.

An interesting contribution of this dissertation is a better understanding of how people scan for answers to questions within documents. To the best of my knowledge this is the first study that showed scanners make loose semantic connections when scanning for particular data within a text document. This highlights the inadequacies

of current systems available to users who employ these systems to read through documents for homework and work-related assignments.

A final contribution of the dissertation is in the areas of participatory design in the design and evaluation of the system interface. This system was both inspired by and designed with the help of users of assistive reading technology. One interesting contribution was the differing feedback from screenreader users versus screen magnifier users. There is a general tendency to group users who are blind and low vision into one category, yet feedback from screenreader users focused largely on getting keystrokes to work properly, while feedback from the screen magnifier consultant was concerned with pop-up boxes and how things enlarged on the screen. Clearly, when designing interfaces, these two groups cannot be considered one.

This dissertation has most notably resulted in an important piece of Assistive Technology. Equally, it has made contributions to interface design, specifically for blind and low-vision readers, and contributions to our knowledge of how visual scanners scan through documents when answering questions.

REFERENCES

[1] Abney, S., Collins, M., and Singhal, A. 2000. Answer Extraction. In Proceedings of ANLP 2000.

[2] Abras, C., Maloney-Krichmar, D., Preece, J. (2004) User-Centered Design. In Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications

[3] Agirre, E., Ansa, O, Hovy, E. and Martinez, D. 2000. Enriching very large ontologies using the www, In ECAI Workshop on Ontology Learning. Berlin, Germany.

[4] Ahmed, F., Borodin, Y., Puzis, Y., and Ramakrishnan, I. V. (2012a). Why Read if You Can Skim: Towards Enabling Faster Screen Reading. In International Cross-Disciplinary ConferenceonWebAccessibility.

[5] Ahmed, F., Borodin, Y., Soviak, A., Islam, M., Ramakrishnan, I. V., and Hedgpeth, T. (2012b). Accessible skimming: faster screen reading of web pages. In Proceedings of the25th annual ACM symposium on User interface software and technology.

[6] Ahmed,F.,Soviak,A.,Borodin,Y.,andRamakrishnan,I.(2013). Non-visual skimming on touch-screen devices. In Proceedings of the 2013 International Conference on Intelligent User Interfaces, IUI '13, pages 435–444, NewYork, NY, USA. ACM.

[7] American Foundation for the Blind: Statistics and Sources for Professionals. 2007. Retrieved on August 7, 2007 from the American Foundation for the Blind Web Site: http://www.afb.org/Section.asp?SectionID=15&DocumentID=1367

[8] Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. 2007. Open information extraction from the Web. In Proceedings of the 20th International Joint Conference on Artificial Intelligence, 2007.

[9] Barker, K., V. K. Chaudhri, S. Y. Chaw, P. Clark, J. Fan, D. Israel, S. Mishra, B. W. Porter, P.Romero, D. Tecuci, and P. Z. Yeh. 2004. A Question-answering system for AP chemistry: Assessing KR&R technologies. In Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference (KR2004), Whistler, Canada, 488-497.

[10] Bigham, J.P., A.C. Cavender, J.T. Brudvik, J.O. Wobbrock, and R.E. Lander, WebinSitu: a comparative analysis of blind and sighted browsing behavior, in Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility. 2007, ACM: Tempe, Arizona, USA.

[11] Bollegala, D., Matsuo, Y., and Ishizuka, M. 2007. Measuring semantic similarity between words using Web search engines. In Proceedings of WWW 2007, 757-766.

[12] Borodin, Y., J.P. Bigham, G. Dausch, and I.V. Ramakrishnan, More than Meets the Eye: A Survey of Screen-Reader Browsing Strategies, in Proceedings of the 19th International World Wide Web Conference, April 26-27, 2010, ACM: Raleigh, USA.

[13] Brill, E., Lin, J., Banko, M., Domais, S. and Ng, A. 2001. Data-Intensive Question Answering. In Proceedings of the TREC-10 Conference, NIST, Gaithersburg, MD, 183-189.

[14] Buchholz, S. 2001. Using Grammatical Relations, Answer Frequencies and the World Wide Web for TREC Question Answering. In Proceedings of the Tenth Text Retrieval Conference (TREC 2001).

[15] Carbonell, J. and Goldstein, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of SIGIR '98, New York, NY, USA 335-336.

[16] Chali, Y. 2002. Generic and query-based text summarization using lexical cohesion. In R. Cohen and B. Spencer (eds), Advances in Artificial Intelligence: 15th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2002, Calgary, Canada, 293-302.

[17] Chen, H., Lin, M, and Wei, Y. 2006. Novel association measures using web search with double checking. In Proceedings of the COLING/ACL 2006. 1009-1016.

[18] Clark, P., C. Fellbaum, J. Hobbs. 2008. Using and Extending WordNet to Support Question-Answering. In Proceedings Fourth Global WordNet Conference (GWC'08), Hungary: University of Szeged, 111-119.

[19] Dyson, M. C., and M. Haselgrove. 2000. The effects of reading speed and reading patterns on the understanding of text read from screen. Journal of Research in Reading, 23. 210-223.

[20] Felbaum, C. 1998. WordNet an Electronic Database, Boston/Cambridge: MIT Press.

[21] Galea, A. 2003. Open-domain Surface-Based Question Answering System. In Proceedings of the Computer Science Annual Workshop (CSAW), University of Malta.

[22] Green, B.F., A.K. Wolf, C. Chomsky and K.Laughery.1961. Baseball:An automatic question answerer. In Proceedings Western Computing Conference, vol. 19, 219-224.

[23] Hearst, M. 1993. TextTiling: A Quantitative Approach to Discourse Segmentation, Technical Report UCB:S2K-93-24.

[24] E. H. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C.-Y. Lin. 2000. Question Answering in Webclopedia. In Proceedings of the TREC-9 Conference. NIST, Gaithersburg, MD. November 2000. 655-664.

[25] Hovy, E.H., U. Hermjakob, and D. Ravichandran. 2002. A Question/Answer Typology with Surface Text Patterns. In Proceedings of the DARPA Human Language Technology Conference (HLT).

[26] Hovy, E. and C.Y. Lin.1997. Automated Text Summarization in SUMMARIST. In *Proceedings of the* Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, 18-24.

[27] Hovy, E.H., C.-Y. Lin, and L. Zhou. 2005. A BE-based Multi-document Summarizer with Sentence Compression**.** *Proceedings of the Multilingual Summarization Evaluation Workshop at the ACL 2005 conference*. Ann Arbor, MI.

[28] Jarvelin. K. and J. Kekalainen. 2002. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20(4), 422–446.

[29] JAWS 18 Documentation.JAWS Screen Reader – Documentation. 2017, Available from: http://www.freedomscientific.com/products/blindness/jawsdocumentation

[30] Katz, B., G. Marton, G. Borchardt, A. Brownell, S. Felshin, D. Loreto, J. L. Rosenberg, B. Lu, F. Mora, S. Stiller, O. Uzuner, and A. Wilcox. 2005. External Knowledge Sources for Question Answering. In Proceedings of the 14th Annual Text REtrieval Conference (TREC2005), November 2005, Gaithersburg, MD.

[31] Kwok, C., O. Etzioni, and D.S. Weld. 2001. Scaling Question Answering to the Web. In Proceedings of the 10th World Wide Web Conference, Hong Kong.

[32] Matsuo, Y., T. Sakaki, K. Uchiyama, and M. Ishizuka. 2006. Graph-based word clustering using Web search engine. In. Proceedings of Empiricial Methods in Natural Language Processing (EMNLP 2006), 542-550.

[33] McLaughlin, G. H.. 1969. Reading at "Impossible" Speeds. Journal of Reading, 12(6):449-454,502-510.

[34] Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013 Efficient Estimation of Word Representations in Vector Space. In: ICLR: Proceeding of the International Conference on Learning Representations Workshop Track, Arizona, USA, 1301–3781.

[35] Moldovan D., Harabagiu, S., Pasca, M., Mihalcea, R., Goodrum, R., Girju, R. and Rus, V. 1999. LASSO: A Tool for Surfing the Answer Net. In Proceedings of the Text Retrieval Conference (TREC-8). November. 563-570.

[36] National Center for Policy Research for Women and Families. 2004. Blind Adults in America: Their Lives and Challenges. February.

[37] National Institute of Child Health and Human Development. 2007. Retrieved on 8/1/2007 from http://www.nichd.nih.gov/publications/pubs/readbro.htm

[38] National Institute for Literacy: Terms. Retrieved 8/1/2007 from the National Institute for Literacy Web Site: http://www.nifl.gov/partnershipforreading/adult_reading/glossary/glossary.html

[39] Norman, D.A. 1988. The Psychology of Everyday Things, New York, Basic Books (Perseus).

[40] Pennington, J., R. Socher, and C. D. Manning. 2014. GloVe: Global vectors for word representation. In Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014). 1532–1543.

[41] Prager, J.M. and J. Chu-Carroll. 2001. Use of WordNet Hypernyms for Answering What-Is Questions. In Proceedings of the TREC-10 Conference, NIST, 309-316.

[42] Raynor, K. 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, Vol. 124, No. 3, 372-422.

[43] Salton, G. and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. In Information Processing & Management, 24 (5): 513-523.

[44] Soricut, R. and E.Brill. 2006. Automatic question answering using the web: Beyond the factoid. Journal of Information Retrieval - Special Issue on Web Information Retrieval, Vol 9, 191–206.

[45] Srihari, R. and W.A. Li. 2000. Question Answering System Supported by Information Extraction. In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-00), 166-172.

[46] Schneiderman, B., and C. Plaisant. Designing the User Interface: Strategies For Effective Human Computer Interaction, 4th edition. 2010 Pearson Education, College Park, Maryland

[47] Takagi, H., S. Kawanaka, and M. Kobayashi, Social Accessibility: achieving accessibility through collaborative metadata authoring, in ASSETS. 2008: Halifax, Canada.

[48] Taylor, S. E. 1962. An evaluation of forty-one trainees who had recently completed the "Reading Dynamics" program. In E.P. Bliesmer & R.C. Staiger (Eds.), Eleventh year book of the National Reading Conference. Milwaukee, Wisc.: The national Reading Conference.

[49] Tullis, T.S. 1986. Optimizing the usability of computer-generated displays. In Proceedings of HCI-86 Conference on People and Computers: Designing for Usability, London, British Computer Society, 604-613.

[50] U.S. Department of Labor: Bureau of Labor and Statistics "Employment Projections" from the Bureau of Labor and Statistics Web Site, http://www.bls.gov/emp/emptab7.htm

[51] Varadarajan, R. and V. Hristidis. 2006. A system for query-specific document summarization, ACM 15th Conference on Information and Knowledge Management (CIKM), Arlington, VA, 622-631.

[52] Voorhees, E.M. 1999. The TREC-8 question answering track report. In Proceedings of the 8th Text Retrieval Conference (TREC-8), Gaithersburg, MD. NIST 77-82.

[53] Wagner, M. and K. Valdes.1995. National Longitudinal Transition Study of Special Education Students, 1987-1991. Menlo Park, CA: SRI International, 1995 [producer], Los Altos, CA: Sociometrics Corporation, 1997 [distributor]. Retrieved on 8/2/2007 from http://www.ciser.cornell.edu/ASPs/search_athena.asp?CODEBOOK=ED-035&IDTITLE=2253

[54] W3C Web Content Accessibility Guidelines (WCAG) 2.0, W3C Recommendation 11 December 2008 from https://www.w3.org/TR/WC

[55] WAI-ARIA. W3C Accessible Rich Internet Applications. 2009 [cited 2009]; Available from: http://www.w3.org/TR/wai-aria.

[56] Wilkinson, S. and S. Payne. 2006. Eye tracking to identify strategies used by readers seeking information from on-line texts. Proceedings of the 13th European Conference on Cognitive Ergonomics, Vol. 250, 115 -116.

[57] Woods, W. 1997. A.Conceptual indexing: "A better way to organize knowledge." Technical Report Sun Microsystems, Inc., SMLITR-97-61.

[58] Zajac, R. 2001. Towards Ontological Question Answering., ACL Open Domain Question Answering Workshop, July 6, 2001, Toulouse.

[59] Zhou, L., C.-Y. Lin, and E.H. Hovy. 2006. Summarizing Answers for Complicated Questions. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC). Genoa, Italy.

Appendix A

QUESTIONS USED DURING SCANNING EXPERIMENTS WITH EYE TRACKING:

Two-page Documents:

1. What effect does China's rising oil prices have on other sectors of its economy?
2. What dietary factors are thought to raise and lower cholesterol?
3. What is responsible for the occurrence of an earthquake?
4. Who created the sport lacrosse and for what reasons?
5. How does marijuana affect the brain?
6. What doubts surround the discovery of this meteorite?
7. Why was Monet's work criticized by the public?
8. According to Piaget, what techniques do children use to adjust to their environment as they grow?
9. Why should you be careful when assuming results of analysis of a small set of data apply to a larger set of data?
10. How do people catch the West Nile Virus?

Longer Documents:

1. Beluga whales seem to summer in distinct areas, yet winter in overlapping areas. Because this may affect the survival of certain populations of Beluga, I need to learn more about the behavior of overlapping populations, specifically in terms of breeding. Why?
2. The author emphasizes meditation’s ability to separate the mental from the physical. Who does he suggest are most constrained by the physical?

Appendix B

EYE TRACKER TEXT FILE

Text file produced by the Tobii Eye Tracking System for one subject scanning for the answer to one question in a 2-page text file. The file specifies the Time, Duration, and in what area of interest (AOI) each hot spot occurred.

Data properties:

Recording date: 5/19/2008
Recording time : 15:40:46:390 (corresponds to time 0)
Study: monet
Subject: al2
Recording: al2wnv14
Screen resolution: 1024 x 768
Coordinate unit: Pixels

Time	Duration	AOI ID	AOI Name	Image/URL
86	439	1	question	WestNileVirus(8)-1.gif
544	159	1	question	WestNileVirus(8)-1.gif
724	498	1	question	WestNileVirus(8)-1.gif
1262	179	1	question	WestNileVirus(8)-1.gif
1462	239	0	Content	WestNileVirus(8)-1.gif
1721	180	1	question	WestNileVirus(8)-1.gif
1920	299	1	question	WestNileVirus(8)-1.gif
2239	100	1	question	WestNileVirus(8)-1.gif
2359	259	1	question	WestNileVirus(8)-1.gif
2638	199	1	question	WestNileVirus(8)-1.gif
2897	299	2	title	WestNileVirus(8)-1.gif
3216	399	3	p1	WestNileVirus(8)-1.gif
3635	498	2	title	WestNileVirus(8)-1.gif
4153	159	3	p1	WestNileVirus(8)-1.gif
4352	239	3	p1	WestNileVirus(8)-1.gif
4611	339	3	p1	WestNileVirus(8)-1.gif
4970	159	3	p1	WestNileVirus(8)-1.gif
5210	140	1	question	WestNileVirus(8)-1.gif
5409	339	1	question	WestNileVirus(8)-1.gif
5768	259	1	question	WestNileVirus(8)-1.gif
6107	239	3	p1	WestNileVirus(8)-1.gif
6366	159	3	p1	WestNileVirus(8)-1.gif
6545	199	2	title	WestNileVirus(8)-1.gif
6765	678	3	p1	WestNileVirus(8)-1.gif
7502	199	3	p1	WestNileVirus(8)-1.gif
7722	758	3	p1	WestNileVirus(8)-1.gif
8539	100	3	p1	WestNileVirus(8)-1.gif
8659	239	3	p1	WestNileVirus(8)-1.gif
8998	219	2	title	WestNileVirus(8)-1.gif
9297	319	2	title	WestNileVirus(8)-1.gif
9636	140	2	title	WestNileVirus(8)-1.gif
9795	199	2	title	WestNileVirus(8)-1.gif
10034	140	3	p1	WestNileVirus(8)-1.gif
10194	339	3	p1	WestNileVirus(8)-1.gif
10593	339	3	p1	WestNileVirus(8)-1.gif
11091	279	3	p1	WestNileVirus(8)-1.gif
11390	399	3	p1	WestNileVirus(8)-1.gif
11829	259	3	p1	WestNileVirus(8)-1.gif
12108	399	3	p1	WestNileVirus(8)-1.gif
12546	159	3	p1	WestNileVirus(8)-1.gif

12726	139	3	p1	WestNileVirus(8)-1.gif
12885	438	3	p1	WestNileVirus(8)-1.gif
13344	120	3	p1	WestNileVirus(8)-1.gif
13503	458	3	p1	WestNileVirus(8)-1.gif
13982	299	3	p1	WestNileVirus(8)-1.gif
14301	159	3	p1	WestNileVirus(8)-1.gif
14480	498	3	p1	WestNileVirus(8)-1.gif
15019	239	3	p1	WestNileVirus(8)-1.gif
15298	179	3	p1	WestNileVirus(8)-1.gif
15497	478	3	p1	WestNileVirus(8)-1.gif
15995	159	3	p1	WestNileVirus(8)-1.gif
16175	179	3	p1	WestNileVirus(8)-1.gif
16374	139	3	p1	WestNileVirus(8)-1.gif
16534	219	3	p1	WestNileVirus(8)-1.gif
16773	299	3	p1	WestNileVirus(8)-1.gif
17092	399	3	p1	WestNileVirus(8)-1.gif
17511	139	3	p1	WestNileVirus(8)-1.gif
17670	219	3	p1	WestNileVirus(8)-1.gif
17909	279	3	p1	WestNileVirus(8)-1.gif
18228	259	3	p1	WestNileVirus(8)-1.gif
18507	498	3	p1	WestNileVirus(8)-1.gif
19026	438	4	p2	WestNileVirus(8)-1.gif
19564	120	4	p2	WestNileVirus(8)-1.gif
19704	339	4	p2	WestNileVirus(8)-1.gif
20062	379	4	p2	WestNileVirus(8)-1.gif
20461	179	4	p2	WestNileVirus(8)-1.gif
20760	199	6	p4	WestNileVirus(8)-1.gif
20980	419	6	p4	WestNileVirus(8)-1.gif
21418	259	7	p5	WestNileVirus(8)-1.gif
21757	199	6	p4	WestNileVirus(8)-1.gif
21976	758	7	p5	WestNileVirus(8)-1.gif
22774	299	7	p5	WestNileVirus(8)-1.gif
23113	140	7	p5	WestNileVirus(8)-1.gif
23272	120	7	p5	WestNileVirus(8)-1.gif
23412	159	7	p5	WestNileVirus(8)-1.gif
23591	140	7	p5	WestNileVirus(8)-1.gif
23751	179	7	p5	WestNileVirus(8)-1.gif
23970	259	5	p3	WestNileVirus(8)-1.gif
24249	139	6	p4	WestNileVirus(8)-1.gif
24488	399	7	p5	WestNileVirus(8)-1.gif
24947	100	7	p5	WestNileVirus(8)-1.gif
25146	179	1	question	WestNileVirus(8)-1.gif
25346	219	0	Content	WestNileVirus(8)-1.gif
25645	578	7	p5	WestNileVirus(8)-1.gif
26243	498	7	p5	WestNileVirus(8)-1.gif
26781	598	7	p5	WestNileVirus(8)-1.gif
27399	319	7	p5	WestNileVirus(8)-1.gif
27738	339	7	p5	WestNileVirus(8)-1.gif
28097	638	7	p5	WestNileVirus(8)-1.gif
28755	159	7	p5	WestNileVirus(8)-1.gif
28934	140	6	p4	WestNileVirus(8)-1.gif
29094	419	6	p4	WestNileVirus(8)-1.gif
29532	319	6	p4	WestNileVirus(8)-1.gif
29871	359	6	p4	WestNileVirus(8)-1.gif
30250	259	5	p3	WestNileVirus(8)-1.gif
30529	120	5	p3	WestNileVirus(8)-1.gif
30669	239	5	p3	WestNileVirus(8)-1.gif
30968	239	5	p3	WestNileVirus(8)-1.gif
31227	139	5	p3	WestNileVirus(8)-1.gif
31446	159	5	p3	WestNileVirus(8)-1.gif
31626	199	5	p3	WestNileVirus(8)-1.gif
31845	219	5	p3	WestNileVirus(8)-1.gif
32084	159	5	p3	WestNileVirus(8)-1.gif
32264	299	5	p3	WestNileVirus(8)-1.gif
32583	160	5	p3	WestNileVirus(8)-1.gif
32762	478	5	p3	WestNileVirus(8)-1.gif
33320	219	6	p4	WestNileVirus(8)-1.gif
33560	239	6	p4	WestNileVirus(8)-1.gif
33839	279	6	p4	WestNileVirus(8)-1.gif
34158	339	6	p4	WestNileVirus(8)-1.gif

34636	259	6	p4	WestNileVirus(8)-1.gif
34915	140	6	p4	WestNileVirus(8)-1.gif
35075	419	6	p4	WestNileVirus(8)-1.gif
35513	179	6	p4	WestNileVirus(8)-1.gif
35713	478	6	p4	WestNileVirus(8)-1.gif
36211	379	6	p4	WestNileVirus(8)-1.gif
36610	299	6	p4	WestNileVirus(8)-1.gif
36949	439	6	p4	WestNileVirus(8)-1.gif
37587	159	6	p4	WestNileVirus(8)-1.gif
37766	598	6	p4	WestNileVirus(8)-1.gif
38384	379	6	p4	WestNileVirus(8)-1.gif
38783	199	6	p4	WestNileVirus(8)-1.gif
39002	479	6	p4	WestNileVirus(8)-1.gif
39501	139	12	p10	WestNileVirus(8)-2.gif
39680	160	10	p8	WestNileVirus(8)-2.gif
39880	159	8	p6	WestNileVirus(8)-2.gif
40059	279	0	Content	WestNileVirus(8)-2.gif
40358	219	0	Blank	
40597	239	0	Content	WestNileVirus(8)-2.gif
40956	279	0	Content	WestNileVirus(8)-2.gif
41255	179	8	p6	WestNileVirus(8)-2.gif
41455	279	8	p6	WestNileVirus(8)-2.gif
41754	140	8	p6	WestNileVirus(8)-2.gif
41913	339	8	p6	WestNileVirus(8)-2.gif
42272	159	8	p6	WestNileVirus(8)-2.gif
42471	957	8	p6	WestNileVirus(8)-2.gif
43448	299	10	p8	WestNileVirus(8)-2.gif
43767	299	10	p8	WestNileVirus(8)-2.gif
44146	419	12	p10	WestNileVirus(8)-2.gif
44644	199	12	p10	WestNileVirus(8)-2.gif
44884	1076	13	p11	WestNileVirus(8)-2.gif
46000	518	13	p11	WestNileVirus(8)-2.gif
46538	797	12	p10	WestNileVirus(8)-2.gif
47356	459	12	p10	WestNileVirus(8)-2.gif
47834	199	12	p10	WestNileVirus(8)-2.gif
48074	299	13	p11	WestNileVirus(8)-2.gif
48393	219	13	p11	WestNileVirus(8)-2.gif
48632	439	13	p11	WestNileVirus(8)-2.gif
49090	319	13	p11	WestNileVirus(8)-2.gif
49509	219	10	p8	WestNileVirus(8)-2.gif
49748	120	10	p8	WestNileVirus(8)-2.gif

Appendix C

SURVEY FOR INTERFACE STUDY

Please complete the following survey questions with the answers that best represent your experience with the system.

Q1

Did you find the system more or less helpful than not using the system in answering the questions?

- ○ A lot less helpful
- ○ Somewhat less helpful
- ○ Neither helpful nor less helpful
- ○ Somewhat more helpful
- ○ A lot more helpful

Q2

Do you think using the system allowed you to answer question more or less quickly than not using the system?

- ○ A lot less quickly
- ○ Somewhat less quickly
- ○ Neither more nor less quickly
- ○ Somewhat more quickly
- ○ A lot more quickly

Q3

Did you use more than one scanning technique in answering the questions (e.g., sentence mode, paragraph mode, ordered sentence mode, and topology mode)

- ○ Yes
- ○ No

Q3b

If you used more than one mode, which modes did you use?

- ☐ Sentence Mode
- ☐ Paragraph Mode
- ☐ Ordered Sentence Mode

- ☐ Topology Mode

Q3c

If you used more than one mode, which mode(s) were most helpful? You may select more than one mode.

- ☐ Sentence Mode
- ☐ Paragraph Mode
- ☐ Ordered Sentence Mode
- ☐ Topology Mode
- ☐ No real difference between helpfulness of the modes

Q4

Did you find the system confusing or straightforward to use?

- ○ Very Confusing
- ○ Somewhat confusing
- ○ Somewhere in the middle
- ○ Mostly straightforward
- ○ Very straightforward

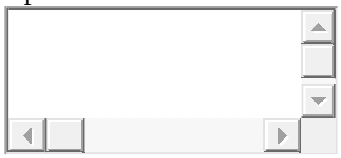
Q5

Did you experience any technical difficulties while using the system?

- ○ Yes
- ○ No

Q5b

If you experienced technical difficulties, could you please describe them.



Q6

Would you use this system again to help find information related to questions in documents?

- ○ Never
- ○ Not Frequently
- ○ Maybe, maybe not
- ○ Occasionally
- ○ Definitely

Q7

Would you recommend this system to a friend?

- Definitely not
- Probably not
- Might or might not
- Probably yes
- Definitely yes

Q19
What assistive technology do you use when reading through a document?

- Screen Reader
- Screen Magnifier
- Other

Q20
Specifically, what assistive technology do you use (e.g., JAWS, MAGic, etc.)?

Q18
Can you suggest keystrokes to replace the existing keystrokes used for moving around the web page (i.e., 1, 2 for sentence mode, 3,4 for paragraph mode, 5,6 for ordered sentence mode, and 7,8 for topology mode)?

Q21
Do you have suggestions for options that would enhance the usability and usefulness of this system?

Q8
Do you have any additional comments and/or suggestions about the system?

Thank you for participating in this experiment. Every effort will be made to update the system based on the feedback you and other participants have given.

Appendix D

RESULTS OF QUESTIONNAIRE:

Question 1

Did you find the system more or less helpful than not using the system

1948	random	somewhat less helpful	2
7227	no	a lot more helpful	5
3505	random	neither less or more	3
5602	no	somewhat more helpful	4
2490	no	somewhat more helpful	4
6060	random	a lot more helpful	5
6090	no	somewhat more helpful	4

Question 2

Do you think using the system allowed you to answer question more or less quickly than not using the system?

1948	random	somewhat less quickly	2
7227	no	a lot more quickly	5
3505	random	neither less or more	3
5602	no	neither more or less	3
2490	no	somewhat more quickly	4
6060	random	somewhat more quickly	4
6090	no	somewhat more quickly	4

Question 3a

Did you use more than one scanning technique in answering the questions?

1948	random	yes
7227	no	yes
3505	random	no
5602	no	yes
2490	no	no
6060	random	yes
6090	no	yes

Question 3b

If you used more than one, which modes did you use?

1948	random	sm,osm
7227	no	pm,osm,tm
3505	random	tm
5602	no	sm,pm,osm
2490	no	sm
6060	random	sm,pm
6090	no	sm,pm

Question 3c

If you used more than one mode, which mode(s) were most helpful?

1948	random	sm
7227	no	no real diff
3505	random	sm
5602	no	
2490	no	sm
6060	random	sm,pm
6090	no	sm

Question 4

Did you find the system confusing or straightforward to use?

1948	random	somewhere in the middle	3
7227	no	very straightforward	5
3505	random	very confusing	1
5602	no	mostly straightforward	4
2490	no	somewhere in the middle	3
6060	random	mostly straightforward	4
6090	no	somewhere in the middle	3

Question 5a

Did you experience any technical difficulties?

1948	random	yes
7227	no	yes
3505	random	no
5602	no	no
2490	no	no
6060	random	no
6090	no	no

Question 5b

If you experienced tech difficulties, could you please describe?

1948	random	I use Jaws 18 and it was unclear at times if the hot keys in the system were actually changing the behavior. I think this was more a function of Jaws than of the system, though.
7227	no	None
3505	random	
5602	no	
2490	no	
6060	random	
6090	no	

Question 6

Would you use this system again to help find info related to questions in documents?

1948	random	definitely	5
7227	no	definitely	5
3505	random	definitely	5
5602	no	occasionally	4
2490	no	maybe, maybe not	3
6060	random	occasionally	4
6090	no	occasionally	4

Question 7

Would you recommend this system to a friend?

1948	random	Probably yes	4
7227	no	Definitely not	1
3505	random	Definitely yes	5
5602	no	Definitely yes	5
2490	no	probably yes	4
6060	random	Probably yes	4
6090	no	Probably yes	4

Question 19

What assistive tech do you use?

1948	random	screenreader
7227	no	Screenreader
3505	random	Other
5602	no	screenreader
2490	no	Screen Magnifier
6060	random	screenreader
6090	no	Screen Magnifier

Question 20

Specifically, what do you use?

1948	random	JAWS
7227	no	NVDA from nv access.org
3505	random	Jaws
5602	no	voice over on MAC
2490	no	zoom
6060	random	JAWS
6090	no	Apple Close view

Question 18

Can you suggest keystrokes to replace existing keystrokes used for moving around the web page?

1948	random	I am not sure but the numbers conflict with standard keys, so if people do not go into the settings to change this as described, or it gets messed up, it won't work. Perhaps have it verbalize something whe switching between modes.
7227	no	
3505	random	no
5602	no	Perhaps use single letters to correspond to the modes; S, O, P and T.
2490	no	none
6060	random	arrow keys they are closer to the edge of the keyboard
6090	no	the numbering was easy, stick with it.

		Question 21
		Do you have suggestions for options that would enhance the usability and usefulness of this system?
1948	random	See above. maybe also have it show the current mode so it could be arrowed over
7227	no	No I don't
3505	random	no
5602	no	I cannot think of any at this time
2490	no	none
6060	random	
6090	no	looks good, keep it up!
		Question 8
		Any additional comments?
1948	random	no
7227	no	No I don't.
3505	random	no
5602	no	While this comment is not related to the system itself, I did find that I tended to default to screen reader navigation commands at first to read the documents and then had to catch myself and not use those. I suppose this more out of habit.
2490	no	
6060	random	
6090	no	Reading is always a challenge as well as dissecting the information. Once I got used to it, It was easier to locate the relative information and move on. A true time saver and with my eye strained induced headaches, a lot less painful. Thank you

Appendix E

IRB APPROVAL FOR EVALUATION OF SKIMMING TECHNIQUES IN QUESTION ANSWERING

HUMAN SUBJECTS PROTOCOL
University of Delaware

Protocol title: Evaluation of Skimming Techniques in Question Answering

Principal Investigator
Name: Debra Yarrington
Contact Phone Number: 302-225-2586
Email address: yarringt@eecis.udel.edu

Advisor (if student PI):
Name: Dr. Kathleen McCoy
Contact Phone Number: 302-831-1956
Email address: mccoy@eecis.udel.edu

Other investigators:

Type of review:
Exempt Expedited Full board

Exemption Category: 1 2 3 4 5 6

Minimal Risk: ✓ yes _____ no

Submission Date: 12/13/07

HSRB Approval Signature	Approval Date
Elizabeth Dugan Peters	2/27/08
HS Number	Approval Next Expires
HS 08-253	2/26/09

Investigator Assurance:
By submitting this protocol, I acknowledge that this project will be conducted in strict accordance with the procedures described. I will not make any modifications to this protocol without prior approval by the HSRB. Should any unanticipated problems involving risk to subjects, including breaches of guaranteed confidentiality occur during this project, I will report such events to the Chair, Human Subjects Review Board immediately.

Signature of Investigator:_____

Date:_____

Appendix F

IRB APPROVAL FOR EVALUATION THE EFFICACY AND USEFULNESS OF A SYSTEM TO ASSIST READERS USING ASSISTIVE TECHNOLOGY IN LOCATING INFORMATION WITHIN A DOCUMENT RELATED TO A QUESTION