

Automated Skimming in Response to Questions for NonVisual Readers

Debra Yarrington

Dept. of Computer and Information Science
University of Delaware
Newark, DE, 19716, USA
yarringt@eecis.udel.edu

Kathleen F. McCoy

Dept. of Computer and Information Science
University of Delaware
Newark, DE, 19716, USA
mccoy@cis.udel.edu

Abstract

This paper presents factors in designing a system for automatically skimming text documents in response to a question. The system will take a potentially complex question and a single document and return a Web page containing links to text related to the question. The goal is that these text areas be those that visual readers would spend the most time on when skimming for the answer to a question. To identify these areas, we had visual readers skim for an answer to a complex question while being tracked by an eye-tracking system. Analysis of these results indicates that text with semantic connections to the question are of interest, but these connections are much looser than can be identified with traditional Question-Answering or Information Retrieval techniques. Instead, we are expanding traditional semantic treatments by using a Web search. The goal of this system is to give non-visual readers information similar to what visual readers get when skimming through a document in response to a question.

1 Introduction

This paper describes semantic considerations in developing a system for giving nonvisual readers information similar to what visual readers glean when skimming through a document in response to a question. Our eventual system will be unique in that it takes both simple and complex questions, will work in an unrestricted domain, will locate answers within a single document, and will return not just an answer to a question, but the information visual skimmers acquire when skimming through a document.

1.1 Goals

Production of our skimming system will require the attainment of three major goals:

1. Achieving an understanding of what information in the document visual skimmers pay attention to when skimming in response to a question
2. Developing Natural Language Processing (NLP) techniques to automatically identify areas of text visual readers focus on as determined in 1.
3. Developing a user interface to be used in conjunction with screen reading software to deliver the visual skimming experience.

In this paper we focus on the first two of these goals. Section 2 will discuss experiments analyzing visual skimmers skimming for answers to questions. Section 3 will discuss developing NLP techniques to replicate the results of Section 2. Section 4 will discuss future work.

1.2 Impetus

The impetus for this system was work done by the author with college students with visual impairments who took significantly longer to complete homework problems than their visually reading counterparts. Students used both ScreenReaders, which read electronic text aloud, and screen magnifiers, which increase the size of text on a screen. While these students were comfortable listening to the screenreader reading at rates of up to 500 words per minute, their experience was quite different from their visual-reading peers. Even after listening to an entire chapter, when they wanted to return to areas of text that contained text relevant to the answer, they had to start listening from the beginning and traverse the document again. Doing

homework was a tedious, time-consuming task which placed these students at a serious disadvantage. It is clear that individuals with visual impairments struggle in terms of education. By developing a system that levels the playing field in at least one area, we may make it easier for at least some individuals to succeed.

2 Visual Skimming

If our intention is to convey to nonvisual readers information similar to what visual readers acquire when skimming for answers to questions, we first must determine what information visual readers get when skimming. For our purposes, we were interested in what text readers focused on in connection to a question. While many systems exist that focus on answering simple, fact-based questions, we were more interested in more complex questions in which the answer could not be found using pattern matching and in which the answer would require at least a few sentences, not necessarily contiguous within a document. From an NLP standpoint, locating longer answers with relevant information occurring in more than one place that may or may not have words or word sequences in common with the question poses an interesting and difficult problem. The problem becomes making semantic connections within any domain that are more loosely associated than the synonyms, hypernyms, hyponyms, etc. provided by WordNet (Felbaum, 1998). Indeed, the questions that students had the most difficulty with were more complex in nature. Thus we needed to find out whether visual skimmers were able to locate text in documents relevant to complex questions and, if so, what connections visual skimmers are making in terms of the text they choose to focus on.

2.1 Task Description

To identify how visual readers skim documents to answer questions, we collected 14 questions obtained from students' homework assignments, along with an accompanying document per question from which the answer could be obtained. The questions chosen were on a wide variety of topics and were complex in nature. An example of a typical question is, "According to Piaget, what techniques do children use to adjust to their environment as they grow?" Documents largely

consisted of plain text, although each had a title on the first page. They held no images and few subtitles or other areas users might find visually interesting. Twelve of the documents were two pages in length, one was eight pages in length, and one was nine pages long. In each case, the answer to the question was judged by the researchers to be found within a single paragraph in the document.

Forty-three visual reading subjects skimmed for the answer to between 6 – 13 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 skimmed 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. Subjects had no time limit for skimming and switched pages by pressing the space bar. When done skimming each document, subjects were asked to select a best answer in multiple choice form (to give them a reason to take the skimming task seriously).

2.2 Results

Results showed that subjects were reliably able to correctly answer the multiple choice question after skimming the document. Of the 510 questions, 423 (about 86%) were answered correctly. The two questions from longer documents 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).

Clearly for the shorter documents, subjects were able to get appropriate information out of the document to successfully answer the question. With that established, we were 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 tracks the path and length of time a subject gazes at a particular point as a subject skims through a document. The system allows us to define Areas of Interest (AOIs) and then track the number of prolonged gaze points

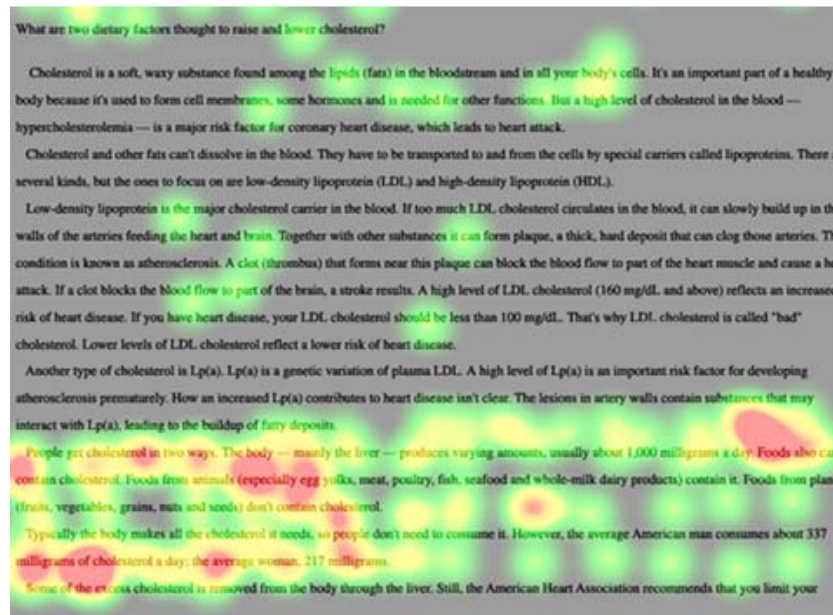


Figure 1. Hot spot image results of skimming for the answer to the question, “What are two dietary factors thought to raise and lower cholesterol?” using the Tobii Eye Tracking System

within those areas of interest. For our analysis, we defined areas of interest as being individual paragraphs. While we purposely chose documents that were predominantly text, each had a title as well. Titles and the few subtitles and lists that occurred in the documents were also defined as separate AOIs. For each skimming activity, the eye tracking system gave us a gaze plot showing the order in which individuals focused on particular areas, and a hot spot image showing the gaze points, with duration indicated with color intensity, that occurred in each AOI (see Figure 1).

In looking at the hot spot images, we found that subjects used three techniques to peruse a document. One technique subjects used was to move their gaze slowly throughout the entire document, indicating that they were most likely reading the document. A second technique used was to move randomly and quickly from top to bottom of the document (described as “fixations distributed in a rough zig-zag down the page” by McLaughlin in reference to speed reading (1969)), without ever focusing on one particular area for a longer period of time. This technique was the least useful to us because it gave very little information. A third technique was a combination of the first two, in which the subject’s gaze darted quickly and randomly around the page, and then appeared to focus on a particular area for an extended period of time.

Figure 1 is a good example of this technique. The data from this group was clearly relevant to our task since their fixation points clearly showed what areas subjects found most interesting while skimming for an answer to a question.

2.3 Analysis of Skimming Data

To determine exactly which AOIs subjects focused on most frequently, we counted the number of gaze points (or focus points) in each AOI (defined as paragraphs, titles, subtitles) across all subjects. In looking at what information individuals focused on while skimming, we found that individuals did focus on the title and subtitles that occurred in the documents. Subjects frequently focused on the first paragraph 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 these users the same information that visual skimmers get when skimming through a document.

We also wanted to look at AOIs that did not have physical features that may have attracted attention. Our conjecture was that these AOIs were focused on by subjects because of their semantic

relationship to the question. Indeed, we did find evidence of this. Results indicated that subjects did focus on the areas of text containing the answer to the question. 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 word it does have in common with the question, ‘West Nile Virus’, is the topic of the document and occurs fairly frequently throughout the document, and thus cannot account for subjects' focusing on this particular paragraph.

The subjects must have made semantic connections between the question and the answer that cannot be explained by simple word matching or even synonyms, hypernyms and hyponyms. In the above example, the ability of the user to locate the answer hinged on their ability to make a connection between the word ‘catch’ in the question and its meaning ‘to be infected by’. Clearly simple keyword matching won't suffice in this case, yet equally clearly subjects successfully identified this paragraph as being relevant to the question. This suggests that when skimming 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 skimmed this document, 15 focused on this paragraph, making it the second most focused on AOI in the document, second only to the paragraph that contained the answer (focused on by 21 of the subjects). The above paragraph occurred within the middle of the second page of the document, with no notable 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.” The paragraph also includes “morose” and “somber” and even “had not been accepted by the world at large.” Subjects appeared to be making a connection between the question topic, Monet's work being criticized by the public, and the above terms. Intuitively, we do seem to make this connection. Yet the connection being made is not straightforward and cannot be replicated using the direct se-

semantic connections that are available via WordNet. 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. Our system must be able to replicate these connections automatically.

Upon further examination, we 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 was even more tenuous. For instance, when skimming for the answer to the question,

"How does marijuana affect the brain?"

the second most frequently focused on paragraph (second to the paragraph with the answer) 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 was this paragraph that subjects chose to focus on. The paragraph is the third to last paragraph on the first page, so its physical location could not explain its attraction to subjects. However, when we looked more closely at the previous paragraphs, we saw 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 skimming through the document would pay more attention to this paragraph because it seems to represent the start of the area that may contain the answer, not to mention conveying topological information about the layout of the document and general content information as well.

Data collected from these experiments suggest that subjects do make and skim for semantic con-

nections. Subjects not only glean information that directly answers the question, but also on content within the document that is semantically related to the question. While physical attributes of text do attract the attention of skimmers, and thus we must include methods for accessing this data as well, it is clear that in order to create a successful skimming device that conveys information similar to what visual skimmers get when skimming for the answer to a question, we must come up with a method for automatically generating loose semantic connections and then using those semantic connections to locate text skimmers considered relevant within the document.

3 NLP Techniques

In order to automatically generate the semantic connections identified above as being those visual skimmers make, we want to explore Natural Language Processing (NLP) techniques.

3.1 Related Research

Potentially relevant methodologies may be found in Open Domain Question Answering Systems. Open Domain Question Answering Systems involve connecting questions within any domain and potential answers. These systems usually do not rely on external knowledge sources and are limited in the amount of ontological information that can be included in the system. The questions are usually fact-based in form (e.g., "How tall is Mt. Everest?"). These systems take a question and query a potentially large set of documents (e.g., the World Wide Web) to find the answer. A common technique is to determine a question type (e.g., "How many ...?" would be classified as 'numerical', whereas "Who was ...?" would be classified as 'person', etc.) and then locate answers of the correct type (Abney et al., 2000; Kwok et al., 2001; Srihari and Li, 2000; Galea, 2003). Questions are also frequently reformulated for pattern matching (e.g., "Who was the first American Astronaut in space?" becomes, "The first American Astronaut in space was" (Kwok et al., 2001; Brill et al., 2002)). Many systems submit multiple queries to a document corpus, relying on redundancy of the answer to handle incorrect answers, poorly constructed answers or documents that don't contain the answer (e.g., Brill et al., 2002; Kwok et al.,

2001). For these queries, systems often include synonyms, hypernyms, hyponyms, etc. in the query terms used for document and text retrieval (Hovy et al., 2000; Katz et al., 2005). In an attempt to answer more complex relational queries, Banko et al. (2007) parsed training data into relational tuples for use in classifying text tagged for part of speech, chunked into noun phrases, and then tagged the relations for probability. Soricut and Brill (2006) trained data on FAQ knowledge bases from the World Wide Web, resulting in approximately 1 million question-answer pairs. This system related potential answers to questions using probability models computed using the FAQ knowledge base.

Another area of research that may lend useful techniques for connecting and retrieving relevant text to a question is query-biased text summarization. With many summarization schemes, a good deal of effort has been placed on identifying the main topic or topics of the document. In query biased text summarization, however, the topic is identified a priori, and the task is to locate relevant text within a document or set of documents. In multidocument summarization systems, redundancy may be indicative of relevance, but should be eliminated from the resulting summary. Thus a concern is measuring relevance versus redundancy (Carbonell and Goldstein, 1998; Hovy et al., 2005; Otterbacher et al., 2006). Like Question Answering systems, many summarization systems simply match the query terms, expanded to include synonyms, hypernyms, hyponyms, etc., to text in the document or documents (Varadarajan and Hristidis, 2006; Chali, 2002)

Our system is unique in that it has as its goal not just to answer a question or create a summary, but to return information visual skimmers glean while skimming through a document. Questions posed to the system will range from simple to complex in nature, and 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. Answering complex questions within a single document when the answer may not

be straightforward in nature poses a challenging problem.

3.2 Baseline Processing

Our baseline system attempted to identify areas of interest by matching against the query in the tradition of Open Domain Question Answering. For our baseline, we used the nonfunction words in each question as our query terms. The terms were weighted with a variant of TF/IDF (Salton and Buckley, 1988) in which terms were weighted by the inverse of the number of paragraphs they occurred in within the document. This weighting scheme was designed to give lower weight to words associated with the document topic and thus conveying less information about relevance to the question. Each query term was matched to text in each paragraph, and paragraphs were ranked for matching using the summation of, for each query term, the number of times it occurred in the paragraph multiplied by its weight.

Results of this baseline ranking were poor. In none of the 14 documents did this method connect the question to the text relevant to the answer. This was expected. This original set of questions was purposely chosen because of the complex relationship between the question and answer text.

Next we expanded the set of query terms to include synonyms, hypernyms, and hyponyms as defined in WordNet (Felbaum, 1998). We included all senses of each word (query term). Irrelevant senses resulted in the inclusion of terms that were no more likely to occur frequently than any other random word, and thus had no effect on the resulting ranking of paragraphs. Again, each of the words in the expanded set of query terms was weighted as described above, and paragraphs were ranked accordingly.

Again, results were poor. Paragraphs ranked highly were no more likely to contain the answer, nor were they likely to be areas focused on by the visual skimmers in our collected skimming data.

Clearly, for complex questions, we need to expand on these basic techniques to replicate the semantic connections individuals make when skimming. As our system must work across a vast array of domains, our system must make these connections “on the fly” without relying on previously defined ontological or other general knowledge. And our system must work quickly: asking

individuals to wait long periods of time while the system creates semantic connections and locates appropriate areas of text would defeat the purpose of a system designed to save its users time.

3.3 Semantically-Related Word Clusters

Our solution is to use the World Wide Web to form clusters of topically-related words, with the topic being the question. The cluster of words will be used as query terms and matched to paragraphs as described above for ranking relevant text.

Using the World Wide Web as our corpus has a number of advantages. Because of the vast number of documents that make up the World Wide Web, we can rely on the redundancy that has proved so useful for Question Answering and Text Summarization systems. By creating the word clusters from documents returned from a search using question words, the words that occur most frequently in the related document text will most likely be related in some way to the question words. Even relatively infrequently occurring word correlations can most likely be found in some document existing on the Web, and thus strangely-phrased questions or questions with odd terms will still most likely bring up some documents that can be used to form a cluster. The Web covers virtually all domains. Somewhere on the Web there is almost certainly an answer to questions on even the most obscure topics. Thus questions containing words unique to uncommon domains or questions containing unusual word senses will return documents with appropriate cluster words. Finally, the Web is constantly being updated. Terms that might not have existed even a year ago will now be found on the Web.

Our approach is to use the nonstop words in a question as query terms for a Web search. The search engine we are using is Google (www.google.com). For each search engine query, Google returns an ranked list of URLs it considers relevant, along with a snippet of text it considers most relevant to the query (usually because of words in the snippet that exactly match the query terms). To create the cluster of words related semantically to the question, we are taking the top 50 URLs, going to their correlating Web page, locating the snippet of text within the page, and creating a cluster of words using a 100-word window surrounding the snippet. We are using only nonstop

words in the cluster, and weighting the words based on their total number of occurrences in the windows. These word clusters, along with the expanded baseline words, are used to locate and rank paragraphs in our question document.

Our approach is similar in spirit to other researchers using the Web to identify semantic relations. Matsuo et al. (2006) looked at the number of hits of each of two words as a single keyword versus the number of hits using both words as keywords to rate the semantic similarity of two words. Chen et al. (2006) used a similar approach to determine the semantic similarity between two words: with a Web search using word P as the query term, they counted the number of times word Q occurred in the snippet of text returned, and vice versa. 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. Sahami and Heilman (2006) used the snippets from a word search to form a set of words weighted using TF/IDF, and then determined the semantic similarity of two keywords by the similarity of two word sets returned in those snippets.

Preliminary results from our approach have been encouraging. For example, with the question, “How does Marijuana affect the brain?”, the expanded set of keywords included, “hippocampus, receptors, THC, memory, neuron”. These words were present in both the paragraph containing the answer and the second-most commonly focused on paragraph in our study. While neither our baseline nor our expanded baseline identified either paragraph as an area of interest, the semantically-related word clusters did.

4 Future Work

This system is a work in progress. There are many facets still under development, including a finer analysis of visual skimming data, a refinement of the ranking system for locating areas of interest within a document, and the development of the system’s user interface.

4.1 Skimming Data Analysis

For our initial analysis, we focused on the length of time users spent gazing at text areas. In future

analysis, we will look at the order of the gaze points to determine exactly where the subjects first gazed before choosing to focus on a particular area. This may give us even more information about the type of semantic connection subjects made before choosing to focus on a particular area. In addition, in our initial analysis, we defined AOIs to be paragraphs. We may want to look at smaller AOIs. For example, with longer paragraphs, the text that actually caught the subject's eye may have occurred only in one portion of the paragraph, yet as the analysis stands now the entire content of the paragraph is considered relevant and thus we are trying to generate semantic relationships between the question and potentially unrelated text. While the system only allows us to define AOIs as rectangular areas (and thus we can't do a sentence-by-sentence analysis), we may wish to define AOIs as small as 2 lines of text to narrow in on exactly where subjects chose to focus.

4.2 Ranking System Refinement

It is worth mentioning that, while a good deal of research has been done on evaluating the goodness of automatically generated text summaries (Mani et al., 2002; Lin and Hovy, 2003; Santos et al., 2004) our system is intended to mimic the actions of skimmers when answering questions, and thus our measure of goodness will be our system's ability to recreate the retrieval of text focused on by our visual skimmers. This gives us a distinct advantage over other systems in measuring goodness, as defining a measure of goodness can prove difficult. In future work, we will be exploring different methods of ranking text such that the system returns results most similar to the results obtained from the visual skimming studies. The system will then be used on other questions and documents and compared to data to be collected of visual skimmers skimming for answers to those questions.

Many variations on the ranking system are possible. These will be explored to find the best matches with our collected visual skimming data. Possibilities include weighting keywords differently according to where they came from (e.g., directly from the question, from the text in retrieved Web pages, from text from a Web page ranked high on the returned URL list or lower, etc.), or considering how a diversity of documents might affect results. For instance, if keywords include

'falcon' and 'hawk' the highest ranking URLs will most likely be related to birds. However, in G.I. Joe, there are two characters, Lieutenant Falcon and General Hawk. To get the less common connection between falcon and hawk and G.I. Joe, one may have to look for diversity in the topics of the returned URLs. Another area to be explored will be the effect of varying the window size surrounding the snippet of text to form the bag of words.

4.3 User Interface

The user interface for our system poses some interesting questions. It is important that the output of the system provide the user with information about (1) document topology, (2) document semantics, and (3) information most relevant to answering the question. At the same time, it is important that using the output be relatively fast. The output of the system is envisioned as a Web page with ranked links at the top pointing to sections of the text likely to be relevant to answering the question.

An important issue that must be explored in depth with potential users of the system is the exact form of the output web page. We need to explore the best method for indicating text areas of interest and the overall topology. The goal is that reading the links simulate what a visual skimmer gets from lightly skimming. The user would actually follow the links that appeared to be "worth reading" in more detail in the same way that skimmers focus in on particular text segments that appear worth reading.

5 Conclusion

This system attempts to correlate NLP techniques for creating semantic connections with the semantic connections individuals make. Using the World Wide Web, we may be able to make those semantic connections across any topic in a reasonable amount of time without any previously defined knowledge. We have ascertained that people can and do make semantic links when skimming for answers to questions, and we are currently exploring the best use of the World Wide Web in replicating those connections. In the long run, we envision a system that is user-friendly to nonvisual and low vision readers that will give them an intelligent way to skim through documents for answers to questions.

References

- S. Abney, M. Collins, and A. Singhal. 2000. Answer Extraction. In *Proceedings of ANLP 2000*, 296-301.
- M. Banko, 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*, 2670-2676.
- D. Bollegala, Y. Matsuo, and M. Ishizuka. 2007. Measuring semantic similarity between words using Web search engines. In *Proceedings of WWW 2007*. 757-766.
- 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.
- J. Carbonell and J. Goldstein. 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.
- Y. Chali. 2002. Generic and query-based text summarization using lexical cohesion, In *Proceedings of the Fifteenth Canadian Conference on Artificial Intelligence*, Calgary, May, 293-303.
- H. Chen, M. Lin, and Y. Wei. 2006. Novel association measures using web search with double checking. In *Proceedings of the COLING/ACL 2006*. 1009-1016.
- C. Felbaum. 1998. *WordNet an Electronic Database*, Boston/Cambridge: MIT Press.
- A. Galea. 2003. Open-domain Surface-Based Question Answering System. In *Proceedings of the Computer Science Annual Workshop (CSAW)*, University of Malta.
- 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.
- E. Hovy and C.Y. Lin. 1997. Automated Text Summarization in SUMMARIST. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, 18-24.
- Boris Katz, Gregory Marton, Gary Borchardt, Alexis Brownell, Sue Felshin, Daniel Loreto, Jesse Louis-Rosenberg, Ben Lu, Federico Mora, Stephan Stiller, Ozlem Uzuner, and Angela Wilcox. 2005. External Knowledge Sources for Question Answering *Proceedings of the 14th Annual Text REtrieval Conference (TREC2005)*, November 2005, Gaithersburg, MD.
- C. Kwok, O. Etzioni, and D.S. Weld. 2001. Scaling Question Answering to the Web. In *Proceedings of the 10th World Wide Web Conference*, Hong Kong, 150-161.
- M. Sahami and T. Heilman. 2006. A Web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of 15th International World Wide Web Conference*. 377-386.
- Chin-Yew Lin and E.H. Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics, In *Proceedings of HLT-NAACL*, 71-78.
- I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim. 2002. SUMMAC: a text summarization evaluation, *Natural Language Engineering*, 8 (1):43-68.
- Y. Matsuo, T. Sakaki, K. Uchiyama, and M. Ishizuka. 2006. Graph-based word clustering using Web search engine. In *Proceedings of EMNLP 2006*, 542-550.
- G. Harry McLaughlin. 1969. Reading at "Impossible" Speeds. *Journal of Reading*, 12(6):449-454,502-510.
- Radu Soricut and Eric Brill. 2006. Automatic question answering using the web: Beyond the factoid. *Journal of Information Retrieval - Special Issue on Web Information Retrieval*, 9:191-206.
- G. Salton, and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24, 5, 513-523.
- E. J. Santos, A. A. Mohamed, and Q. Zhao. 2004. "Automatic Evaluation of Summaries Using Document Graphs," Text Summarization Branches Out. *Proceedings of the ACL-04 Workshop*, Barcelona, Spain, 66-73.
- R. Srihari 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.
- R. Varadarajan and V. Hristidis. 2006. A system for query-specific document summarization, *ACM 15th Conference on Information and Knowledge Management (CIKM)*, Arlington, VA, 622-631.