Access to Multimodal Articles from Popular Media for Individuals with Sight Impairments

Sandra Carberry, University of Delaware
Stephanie Elzer, Millersville University
Kathleen McCoy, University of Delaware
Seniz Demir, Scientific and Technological Research Council of Turkey
Peng Wu, University of Delaware
Charles Greenbacker, University of Delaware
Daniel Chester, University of Delaware
Edward Schwartz, Carnegie Mellon University
David Oliver, Millersville University
Priscilla Moraes, University of Delaware

Although intelligent interactive systems have been the focus of many research efforts, very few have addressed systems for individuals with disabilities. This article presents our methodology for an intelligent interactive system that provides individuals with sight impairments with access to the content of information graphics (such as bar charts and line graphs) in popular media. The article describes the methodology underlying the system's intelligent behavior, its interface for interacting with users, examples processed by the implemented system, evaluation studies of the methodology, and a preliminary evaluation of the effectiveness of the overall system. This research advances universal access to electronic documents.

Additional Key Words and Phrases: Human-computer interaction, intelligent systems, multimodal, accessibility, blind individuals

ACM Reference Format:
Carberry, S., Elzer, S., McCoy, K., Demir, S., Wu, P., Greenbacker, C., Chester, D., Schwartz, E., Oliver, D. 2011, Moraes, P. 2012. Access to Multimodal Articles for Individuals with Sight Impairments. ACM V, N, Article A (January YYYY), 48 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

Although intelligent interactive systems have been the focus of many research efforts, only a few projects have addressed systems for individuals with disabilities. ICICLE [Michaud and McCoy 2006] is an interactive system that identifies errors in written English by students whose first language is American Sign Language, hypothesizes the cause of those errors, and provides feedback to the students on writing

This paper is based upon work supported by the National Institute on Disability And Rehabilitation Research under grant H133G080047 and by the National Science Foundation under grants IIS-0534948 and III-1016916.

Author's addresses: S. Carberry and K. McCoy and C. Greenbacker and D. Chester and P. Moraes, Computer Science Department, University of Delaware; S. Demir, Center of Research for Advanced Technologies of Informatics and Information Security, Scientific and Technological Research Council of Turkey; S. Elzer and D. Oliver, Computer Science Department, Millersville University; P. Wu, Microsoft; E. Schwartz, Electrical and Computer Engineering Department, Carnegie Mellon University.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

correct English. Several lines of research have led to improved methods of augmentative and alternative communication for individuals with communicative impairments [Trnka et al. 2009; Gips 1998; Alm et al. 1987; Waller et al. 1992; Newell et al. 1992; MacAulay et al. 2002; Waller et al. 2001; Todman and Alm 2003; Black et al. 2010]. Other work, such as that of Ferres [Ferres et al. 2010], provides access to information graphics (non-pictorial graphics such as bar charts, line graphs, etc.) for individuals with sight impairments; however, although bar charts are treated differently from line graphs, their system provides the same information for each instance of a graph type (i.e., all summaries of line graphs contain the same sorts of information). They do not involve reasoning about the content of the graphic and thus do not rise to the level of an *"intelligent"* system.

We have developed an intelligent, interactive system (SIGHT) for providing individuals with sight impairments with access to the content of information graphics from popular media. SIGHT is *interactive* in that it provides the user with the ability to request a summary of a graphic at a relevant point in a document and offers the opportunity to request follow-up information if desired. This summary is presented as text in a dialog window which can then be read to the user with screen reader software or read by the user with a screen magnifier tool. SIGHT exhibits *intelligence* in several ways. These include:

— offering the user access to an information graphic at the point in the text that is most relevant to the graphic
— hypothesizing the high-level intended message conveyed by a bar chart or a line graph by reasoning about communicative signals selected by the graph designer
— using heuristics based on a corpus study to identify what is being measured on the dependent axis (which often is not explicitly labelled) and to realize it as text
— applying content identification rules based on human subject experiments to select propositions for inclusion in an initial summary of a graphic and subsequent responses
— applying metrics that take into account sentence complexity in aggregating propositions into sentences

This paper describes the interactivity of the SIGHT system and presents the methodology underlying its intelligent behavior. Section 2 motivates the need for access to information graphics, the overall approach that we have taken, and discusses related work on increasing the accessibility of graphics. Section 3 gives an overview of SIGHT's user interface and architecture, and Section 6 expands on this to present our method for identifying where in a document to offer access to an information graphic rather than merely where it is encountered when reading the document. Section 4 presents our Bayesian system for hypothesizing the intended message of an information graphic; we hypothesize that a graphic's intended message should form the core of its initial summary. Unfortunately, graphics often do not explicitly label the dependent axis with what is being measured; Section 5 describes our methodology for identifying an appropriate referent. Section 7 describes our approach for selecting additional salient propositions to include in the initial summary, and Section 8 gives a brief overview of how the selected propositions are structured into a coherent presentation. Section 9 describes our approach to providing further information about the graphic when the initial summary did not satisfy the user's needs. Section 10 presents our preliminary evaluation studies which illustrate the success of the SIGHT system. Although we believe that our methodology is applicable to most types of graphs that appear in popular media, the implementations presented in this paper are limited to simple bar charts and line graphs. We chose simple bar charts and line graphs since

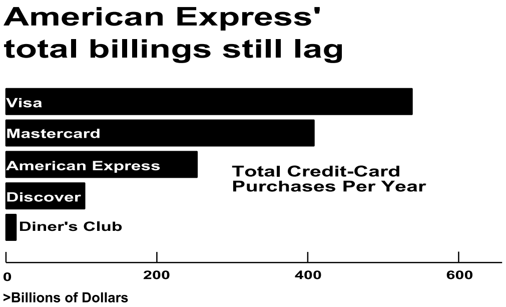


Fig. 1: A graphic from Businessweek.

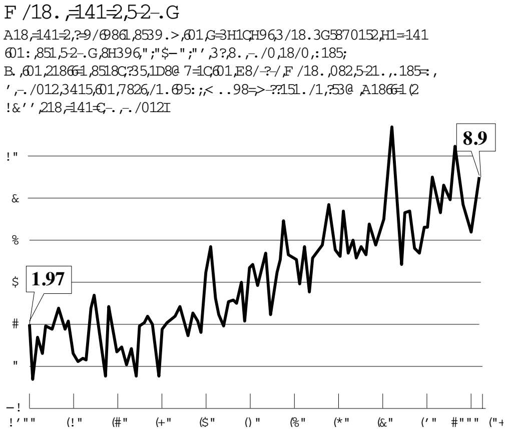


Fig. 2: A graphic from USA Today.

they are two very different kinds of graphs (discrete versus continuous) and the messages conveyed by these two kinds of graphs represent sufficient variability to demonstrate the efficacy of our methodology. Our current work is addressing grouped bar charts [Burns et al. 2012], and in the future we will consider multiple line graphs, pie charts, and composite graphs comprised of more than one graph type. Throughout the paper, we will use examples from our implemented system to illustrate how SIGHT processes a document and its information graphics and provides access to a graphic's content.

2. MOTIVATION

Information graphics in popular media generally have a communicative goal or message that they are intended to convey. For example, the graphic in Figure 1 ostensibly is intended to convey that American Express ranks third in total credit card purchases each year, and the graphic in Figure 2 is intended to convey that there has been a changing trend in ocean levels (relatively stable from approximately 1900 to around 1930 and then rising from around 1930 to 2003). We contend that this intended message captures the high-level content of the graphic and its primary contribution to the multimodal document in which it appears.

In contrast with scientific documents which explicitly refer to and explain their information graphics, the text of articles from popular media often do not capture the content of their information graphics [Carberry et al. 2006]. Moreover, captions on graphics often do not capture the graphic's high-level content [Elzer et al. 2005], as shown by the graphs in Figures 1 and 2. Instead, the reader is expected to assimilate the information graphics and integrate their communicative goals into the communicative goals of the text in order to fully comprehend the document. Thus information graphics in popular media cannot be ignored.

Our research has the goal of providing blind individuals with effective access to documents containing information graphics in popular media. Instead of rendering the graphic in an alternative medium (such as via sound, touch, or a detailed description of what the graph looks like), our approach is to provide the user with the high-level content of the graphic along with an interactive facility for requesting follow-up responses providing more detailed information about the graphic. Thus our system addresses the needs of individuals who want just the high-level contribution of the graphic to the article that they are reading as well as the needs of individuals who want more in-depth knowledge about the graphic's content.

The population of people who are blind or visually impaired is extremely diverse. Differences stem from, among other things, severity and type of vision loss, age of onset (from birth to elderly), underlying causes of vision loss, other disabilities the individual may have, assistive technologies used (e.g., screenreaders, screen magnifiers), as well as education and life experiences. The experiences with and ability to reason about information graphics cannot be determined by any simple combination of these factors. For instance, a person who becomes blind from an early age may have never seen an information graphic, but may have been educated in science and learned to interpret scientific graphics rendered with sounds or given in tactile representations. Such an individual may have a very good grasp on how to interpret graphics (and the vocabulary used to describe them).

In part to reach the widest population, the SIGHT system does not describe the graphic using graphical terms (something that would not be helpful for someone who has not worked with them), but rather focuses on the information content that the author decided to render in graphical form. Though the system does identify an information graphic by type (e.g., bar chart, line graph), the textual description does not require graphical reasoning ability or familiarity to interpret. As is mentioned in Section 9, in some of our studies we found that people who were congenitally blind, and without familiarity with information graphics, had no idea what additional information could be gleaned from a graphic beyond the initial summary we provided. Thus they did not know what kind of follow-up questions would be appropriate. This was a major reason for making the follow-up question interface menu-driven.

2.1. Related Work

In supporting the needs of individuals who are seeking just the high-level contribution of the graphic to the article as well as those individuals who desire more in-depth knowledge about the graphic's content, our work can be viewed as an interesting variant of some of the core problems in information visualization. In information visualization, the issues of overview+detail and focus+context arise from the competing needs of the user and limited viewing area [Chen 2004; Cockburn et al. 2008]. The user might want very detailed visual data, but still require high-level information in order to keep the data in context [Chen 2004; Cockburn et al. 2008]. In information visualization, the challenge is finding intuitive and novel ways to display the information so that the viewing space is utilized as efficiently as possible. In our work, the challenge is designing an intuitive interface that allows a sight-impaired user to obtain the categories and levels of detail of information that they want in appropriately-sized cognitive chunks.

A great deal of research has been concerned with providing people who are blind with access to graphical information, with primary emphasis on scientific data exploration graphs [Lewandowsky and Spence 1989]. The task has been to render the graphic for "viewing" in an alternative modality (such as touch or sound) or to convey a depiction of the graph (labels, values, caption, etc.) via speech. For instance, researchers have investigated techniques for rendering visual information in either auditory (non-speech sounds or tones) [Meijer 1992; Flowers and Hauer 1995; Roth et al. 2002; Brown and Brewster 2003; Alty and Rigas 2005; Cohen et al. 2005; McGookin and Brewster 2006] or haptic/tactile (raised dots or lines, force feedback devices) [Ina 1996; Fritz and Barner 1999; Krufka and Barner 2006; Jayant et al. 2007; Goncu and Marriott 2008], or a combination of these [Kennel 1996; Ramloll et al. 2000; Yu and Brewster 2002; 2003; Yu et al. 2003; Goncu et al. 2010]. Visually important cues in images have been mapped to appropriate tactile analogues [Nayak and Barner 2004; Way and Barner 1997a; 1997b]. These approaches represent a direct translation of visual information into an alternate modality; they do not attempt to capture the meaning of the content. Kurze [Kurze 1995] uses text to convey the content of a graphic — in this case, a verbal description of the diagram's properties (e.g., the style of the diagram, the number of data sets, the labels, ranges of axes). While such approaches are extremely valuable for graphics in scientific articles, the interpretation of these graphs requires training [Walker and Nees 2005], imposes a great cognitive load, may require special hardware and preparation of the graphic itself, and the approaches fail to capture the intention of graphics in popular media.

Some researchers have concentrated their efforts on communicative graphs and providing access in text [Ferres et al. 2007; Ferres et al. 2010] along with inexpensive haptics [Abu Doush et al. 2009]. For example, the iGRAPH-Lite system [Ferres et al. 2007; Ferres et al. 2010] is an interactive system designed specifically for graphics that appear in Statistics Canada's *The Daily*. In addition to a textual summary, iGRAPH-Lite allows the user to access, via keyboard commands, the low-level values of the graphic. This research does make different summaries on the basis of graph type. However, it does not make distinctions between different communicative intentions. Thus, for example, summaries of two different bar charts would include the same information even though they might have very different underlying intentions.

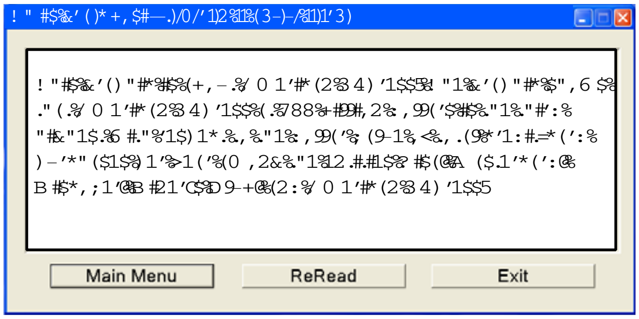
These research efforts differ significantly from our approach, which uses reasoning about the communicative signals in the graph to identify its intended message and provide the user with a summary of the high-level knowledge conveyed by the graphic. Thus our system exhibits a level of *intelligent* behavior that is absent from the other systems.

3. SIGHT: AN INTERACTIVE SYSTEM FOR BLIND INDIVIDUALS

3.1. Interface Design

We have designed and implemented an interactive system called SIGHT for providing blind individuals with effective access to the content of information graphics in popular media. An important design consideration in implementing the SIGHT system was to make the system work as seamlessly as possible within the existing computing environment of most visually impaired users, while allowing them to access the content of information graphics available on the web. Many visually impaired users utilize *screen reading software* to interpret what is on the screen and send that content to a speech synthesizer. The most commonly used screen reader worldwide is JAWS by Freedom Scientific, running on Windows operating systems using the Internet Explorer web browser. Therefore, the SIGHT system has been implemented and tested on a Windows 7 system using JAWS version 11 and Internet Explorer 8.

Just as sighted users develop particular habits in interacting with software, so do visually impaired users. It was our intention in developing SIGHT to preserve users' customary browsing habits as much as possible, since fewer changes in users' habits will result in their ability to learn the system more quickly, and they will be more receptive to using the system on a regular basis. Research done by the American Foundation for the Blind (AFB) [Gerber 2002] showed two basic patterns for sight impaired users when navigating web pages: 1) *scrollers* would rather have the entire screen read to them before choosing a navigational button or action, while 2) *searchers* like to find a particular piece of information and then move on. As a screen reader, JAWS supports both scrolling and searching as patterns of interacting with web pages. Scrollers can simply listen to the content of the web page as JAWS interprets it, taking action when they are ready. Searchers can use the TAB or arrow keys to navigate more quickly through content, taking action as soon as they locate what they are looking for.



```
! " #$%&' ()* +, $#—.)/0/'1)2%1%(3–)–/%1)1'3)
```

```
! "#$%&'() "#*%#$%(+, – .%/ 0 1'#* (2%3 4) '1$$5%! "1%&'() "#*%$", 6 $%
." ( .%/ 0 1'#* (2%3 4) '1$$%( .%788%+#99#, 2%: , 99('$%#$%."1%."#':%
"#&."1$.%6 #."%'1$) 1* .%., %."1%: , 99('%; (9–1%, <%., .(9%*'1: #.=* (':%
) – '*" ($1$%) 1'%>1 ('%(0 , 2&%."1%12 .#.#1$%? #$(@%A ($.1'* (':@%
B #$*, ; 1'@%B #21'C$%D 9–+@%(2: %/ 0 1'#* (2%3 4) '1$$5
```

Fig. 3: Initial summary.

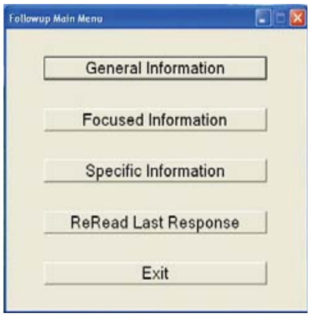


Fig. 4: Main follow-up menu.

Ideally, if the developer of the web page being accessed has supplied alternative text (or “alt text”) for graphics in the html, a screen reader (such as JAWS) will read this text to the user when the graphic comes into focus in the screen reader. However, the vast majority of web pages are developed without broad accessibility in mind, and alt text is not supplied, thus making the content of the document’s graphics inaccessible to a visually impaired user. SIGHT is designed to naturally supplement the processing of alt text if it is present and to replace it if it is not. When a web page is opened in Internet Explorer, the SIGHT system performs a scan of all of the graphics on that page, performing rudimentary image processing to determine if the image is likely to contain an information graphic (currently, either a simple bar chart or a single-line line graph). If an image appears to be an information graphic, the SIGHT system appends a message to the existing alt text (if any) informing the user that CONTROL+Z can be pressed to launch the SIGHT system in order to obtain information about the graphic’s content. By inserting text into the alt text, the SIGHT system is working naturally with the tools that are familiar to visually impaired users, and they will hear the alt text when they encounter the information graphic.

When an information graphic is in focus for the user and CONTROL+Z[1] has been pressed, the SIGHT system begins processing the information graphic. A Visual Extraction Module (VEM) analyzes the graphic and produces an XML representation containing information about the components of the information graphic including the graphic type (bar chart, line graph, etc.) and the caption of the graphic. For a bar chart, the representation includes the number of bars in the graph, the labels of the axes, the caption, and information for each bar such as the label, the height of the bar, the color of the bar, and so forth. For a line graph, the representation includes the beginning and end points of each straight line segment, the labels of the axes, any annotations on the points in the line graph along with their location, the caption, and so forth. The XML representation contains all the detail necessary to redraw the graphic. Although the VEM must process a raw image, the task is much more constrained, and thus much easier, than most image recognition problems. Currently, the VEM can handle electronic images of simple bar charts and line graphs that are clearly drawn in a fixed set of fonts and with standard placement of labels and captions. Current work is removing these limitations. This paper will not be concerned further with the VEM.

When the system begins processing, a message will be presented in a dialog window stating that processing has begun, so that the screen reader can read the message. This immediate feedback is important to a visually impaired user as they have no other way of knowing if they successfully activated processing of the information graphic. The

[1] This key combination was chosen so as not to conflict with existing navigation commands in JAWS.

next feedback that the user will receive from the SIGHT system is an initial summary containing the underlying message of the graphic. This will be conveyed via a new window that will appear in the focus of the screen reader and be read by JAWS (e.g., the dialog window shown in Figure 3 for the graphic in Figure 1). The title of the window will be the caption of the information graphic, the contents of the window will be the summary. Although the content of all dialog windows will be read to the user by JAWS, users who use a screen magnifier can read the system responses and see the windows that are displayed. 16-point or larger fonts are used throughout the SIGHT system so that the text is not overly distorted by magnification [Coyne and Nielsen 2001]. All dialog windows in the system present a set of navigational menu options, all of which are associated with unique keyboard accelerators[2]. For example, there are three buttons in the initial summary dialog window: Main Menu, ReRead and Exit. The "ReRead" button (associated with "Shift+R" keystroke combination) is an important element of the interface since the ability to repeat information when it has not been properly heard the first time is critical for visually impaired users. The Exit button (associated with "Shift+E" keystroke combination) allows the user to exit the SIGHT system if the initial summary provided sufficient information about the information graphic. The Main Menu button(associated with "Shift+M" keystroke combination) provides the interface to the rest of the SIGHT system via a main follow-up window (as shown in Figure 4). From that dialog window, the user can request general, focused or specific follow-up information to what they have already learned about the information graphic. *General Follow-up* is a request for more knowledge about the graphic without specifying any particular kind of knowledge (similar to saying *"Tell me more"*), *Focused Follow-up* is a request for further information about some particular aspect of the graphic such as its trends, and *Specific Follow-up* is a request for detailed information about some entity in the graphic such as a particular bar. Follow-up responses are discussed further in Section 9.

3.2. System Architecture

Figure 5 presents an overview of the architecture of the SIGHT system. A web page is input to the *User Interface* which handles communication with the user. As discussed in Section 3.1, the web page is read to the user via JAWS screenreading software. User requests, whether for an initial summary of an information graphic or for followup information about the graphic, are sent to the *Interaction Module* which is responsible for invoking the appropriate submodules. If the request is for an initial summary of the graphic, the *Interaction Module* sends the electronic image of the graphic to the *Visual Extraction Module* which produces an XML representation of the graphic; this XML representation is sent to the *Intended Message Recognition Module* (described in Sections 4 and 5) that infers the graphic's intended message (which will serve as the core of the initial summary) and returns it to the Interaction Module. For both requests for initial summaries and requests for followup information, the Interaction Module invokes the *Generation Module* which is responsible for identifying the appropriate content of the response (*Content Identification* described in Sections 7 and 9.1), coherently structuring and organizing it (*Text Structuring and Aggregation* and *Sentence Ordering* described in Section 8), and realizing it in natural language (*Sentence Generation*). The English language response is returned to the *Interaction Module* which

[2] Users who use a screenreader can tab to move through the menu options, and press the space bar or the associated keyboard accelerator in order to select the active menu option. In such cases, the options would be read to the user by JAWS. The menu options can also be activated by clicking its corresponding button; this can be done by users with low-vision who use screen magnifiers.

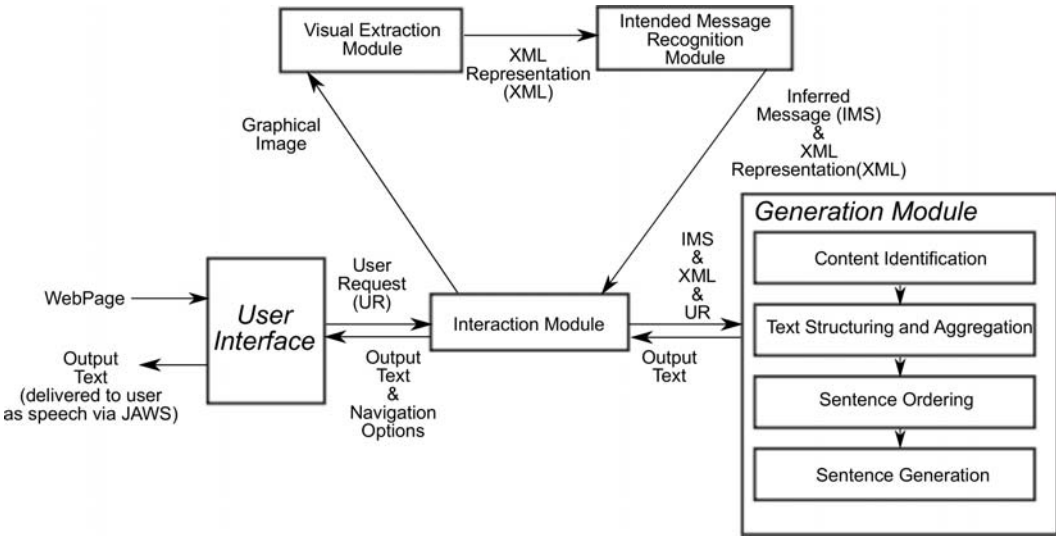


Fig. 5: Architecture of the SIGHT system

returns it along with available navigation options to the *User Interface*; the *User Interface* then delivers the response to the user via speech.

4. RECOGNIZING AN INFORMATION GRAPHIC'S INTENDED MESSAGE

Information graphics in popular media generally have a communicative goal or intended message. As argued by Clark [Clark 1996], language is more than utterances; it is **any action** that is intended to convey a message, or lack of an action when one is expected. For example, pointing, a deliberate head nod, rolling of the eyes, or even failure to respond to an invitation (known to have been received) each convey a message. Thus, adopting Clark's view that communication is not limited to verbal behavior, we view an information graphic as a communicative action, and we exploit an information graphic's communicative signals, such as coloring one part of the graphic differently from others, as evidence in a Bayesian network that hypothesizes the graphic's intended message. This intended message constitutes the most important high-level knowledge that should be included in an initial summary of the graphic.

Figure 6 presents an overview of the *Intended Message Recognition Module* that is part of the SIGHT architecture displayed in Figure 5. If the graph is a line graph, it is first segmented into visually distinguishable trends (described in Section 4.2.1) and the XML representation is augmented to capture the segmentation. For both line graphs and bar charts, a set of candidate intended messages are constructed (described in Section 4.2) and evidence that may signal the intended message is extracted from the graphic (described in Section 4.1). The candidate messages and extracted evidence are used to build the Bayesian network (described in Section 4.3) that hypothesizes the graphic's intended message.

4.1. Communicative Signals in Information Graphics

We have identified several different kinds of communicative signals that can appear in information graphics and serve as evidence of the communicative intention. These can roughly be classified as perceptual task effort, salience devices, and message category cues. We extract any such cues from an information graphic, and use their presence or

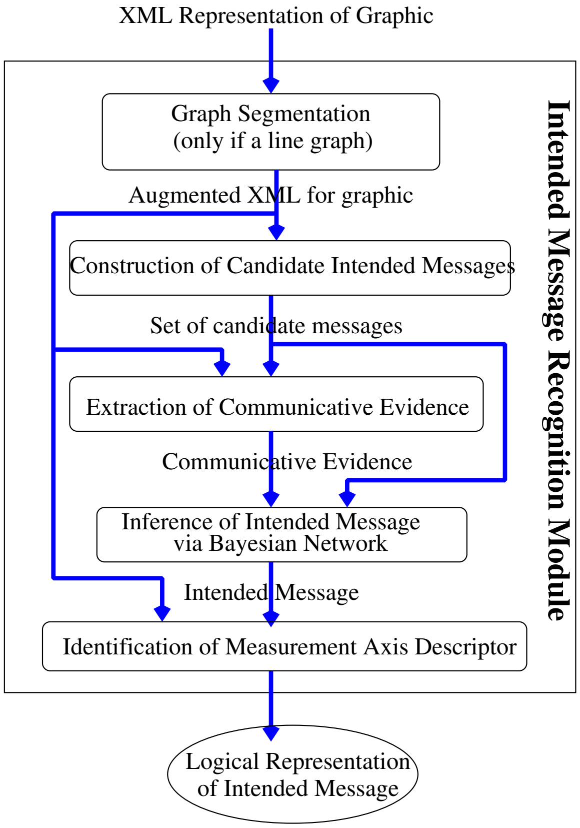

XML Representation of Graphic
Intended Message Recognition Module
Graph Segmentation (only if a line graph)
Augmented XML for graphic
Construction of Candidate Intended Messages
Set of candidate messages
Extraction of Communicative Evidence
Communicative Evidence
Inference of Intended Message via Bayesian Network
Intended Message
Identification of Measurement Axis Descriptor
Logical Representation of Intended Message

Fig. 6: Architecture of the Intended Recognition Module

absence as evidence in a Bayesian network that hypothesizes the graphic's intended message.

AutoBrief [Green et al. 2004] is a system for intelligent multimedia generation. One of their hypotheses was that a graphic designer attempts to design a graphic that best enables the viewer to perform the perceptual and cognitive tasks necessary for identifying the graph's intended message. Thus we hypothesize that the relative difficulty of different perceptual tasks serves as a communicative signal about which tasks the viewer was intended to perform. Consider, for example, the graphics in Figure 7. The message conveyed by the graphic in Figure 7a (and in fact the message intended by the authors of the report[4]) is that the salary of females is consistently less than that of males for each of the science and engineering disciplines. The graphic in Figure 7b depicts the same data, but the organization of the graphic does not facilitate an easy comparison of male versus female salaries in each discipline. Thus it does not convey the same high-level message as the graphic in Figure 7a, although the same information *could* be extracted from both graphics.[5] As noted by [Larkin and Simon 1987],

[3] The leftmost graphic appeared in the 2000 Report of the NSF Committee on Equal Opportunities in Science and Engineering [NSF Committee on Equal Opportunities in Science and Engineering 2000].

[4] We know the intended message since a colleague served on the NSF panel that prepared the report.

[5] The intended message of the graph on the right side of Figure 7 might be that both male and female salaries are lowest in the life and social sciences and highest in engineering.

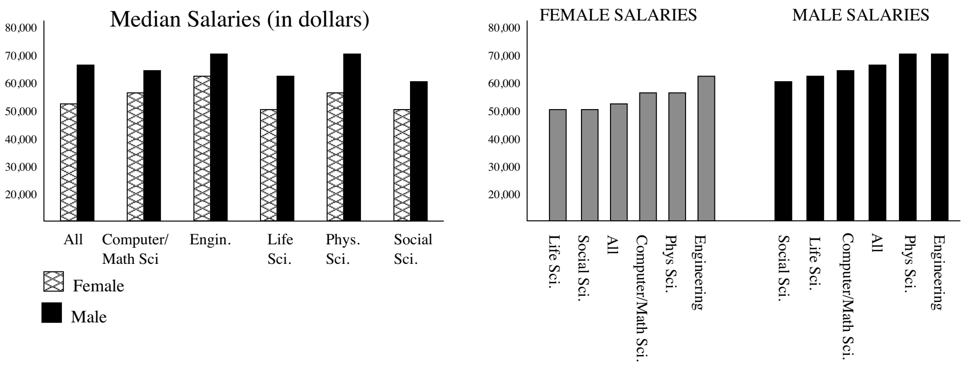


Fig. 7: Two Alternative Graphs from the Same Data.[3]

information graphics that are *informationally* equivalent (all of the information in one graphic can also be inferred from the other) are not necessarily *computationally* equivalent.

Thus we treat relative perceptual task effort as a communicative signal in bar charts; since line graphs do not consist of discrete entities such as bars, perceptual effort does not appear to play the same significant role in line graphs and thus is not considered there. For bar charts, we constructed a set of rules, represented as condition-computation pairs, that estimate the perceptual task effort for different perceptual tasks on bar charts. These rules are based on research by cognitive psychologists and were validated with eye-tracking experiments. For example, the rule shown in Figure 8 estimates the perceptual effort required to obtain the exact value for an attribute in a bar chart given that the user is already focused on the bar. Rule-B1 captures a situation in which the bar is annotated with its value; the estimate is 150 units for discriminating the label (based on work by [Lohse 1993]) and 300 units for recognizing a 6-letter word (based on work by [John and Newell 1990]). Rule-B2 captures the situation in which there is no annotated value but the top of the bar is aligned with a label on the y-axis. A separate rule captures the cognitive effort required if the viewer must extrapolate between two labels on the y-axis. [Elzer et al. 2006] provides details about the rules and the eyetracking experiments that validated them.

```
Rule-B1:Estimate effort for task
          PerceiveValue(<viewer>, <g>, <att>, <t>, <v>)

Graphic-type: bar-chart

Gloss: Compute effort for finding the exact value <v> for attribute <att>
          represented by top <t> of a bar <b> in graph <g>

Conditions:
     B1-1: IF the top <t> of bar <b> is annotated with a value,
             THEN effort=150 + 300
     B1-2: IF the top <t> of bar <b> aligns with a labeled tick mark on
             the dependent axis, THEN effort=230 + (scan + 150 + 300) x 2
```

Fig. 8: Rule for Estimating Effort for the Perceptual Task *PerceiveValue*.

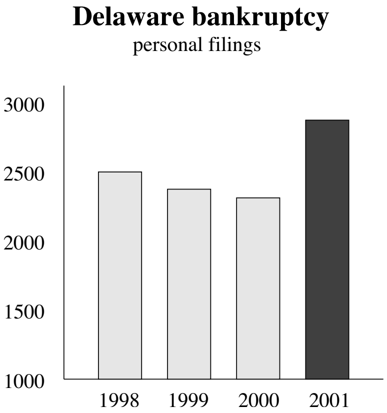


Fig. 9: A bar chart from The Wilmington News Journal.

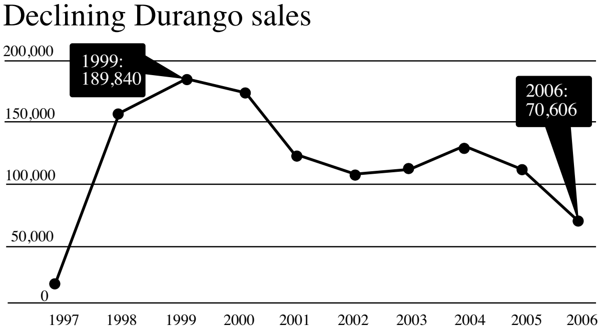


Fig. 10: A line graph from The Wilmington News Journal.

A second class of communicative signal is attention-getting devices that make an entity salient in a graphic. These include a graph designer choosing to color a bar differently from other bars in a bar chart, to annotate an entity in a graphic with its value, or to include a referent to an entity in the graph's caption. For example, the bar for 2001 is colored differently in the graphic in Figure 9, two of the points are annotated in the graphic in Figure 10, and a referent to the bar labeled *American Express* is included in the caption for the graph in Figure 1. Such design decisions make the entity salient and suggest that it plays a prominent role in the graphic's intended message — that is, as a parameter in the logical representation of the graphic's message. The XML representation produced by the Visual Extraction Module captures annotations, coloring, captions, etc. Thus our system extracts from the XML representation of the graph any salience signals present in the graphic.

A third class of communicative signal is cues in the caption that suggest a particular category of intended message. For example, the verb *declining* in the caption of Figure 10 is such a cue; it might suggest a message about a declining trend or a message about a changing trend where the change is from rising or stable to declining. We identified a set of such cue verbs and organized them into semantic similarity categories [Elzer 2005]; for example, the verbs *rise* and *soar* reside in one category whereas the verbs *rebound* and *recover* reside in a different category. The presence of a verb in one of these categories in the caption (or a noun or adjective derived from one of these verbs) is treated as a communicative signal. Our system uses a part-of-speech tagger [TreTagger] and a stemmer [Porter] to extract message-cue signals from the caption.

Each of these signals is treated independently and entered into a Bayesian network (see Section 4.3) that hypothesizes the graphic's intended message. We performed a study on bar charts to assess the impact of each kind of communicative signal on recognizing the correct intended message [Carberry and Elzer 2007]. That study showed that perceptual task effort was the communicative signal that most affected system success. However, the study also showed that the evidence sources compensate for one another: when one source of evidence is disabled, cues from other sources generally provide evidence that often enables recognition of the intended message.

4.2. Constructing Candidate Intended Messages

In order to infer the message intended by an information graphic, we use a Bayesian network. For each new graph, the Bayesian network must consider possible intended messages and use the graphic's communicative signals to arbitrate among the candidate messages. This requires that we construct candidate messages that the system will arbitrate among. Research into graphic design [Tufte 1983] and our analysis has shown that given a graph type (e.g., line graph, bar chart) only a limited number of kinds of intended messages are possible. We analyzed a corpus of simple bar charts and a corpus of line graphs and identified 12 categories of messages for bar charts and 10 for line graphs. Both include categories such as *Rising-trend* and *Change-trend*. However, the categories for bar charts include *Rank* (conveying the rank of an entity represented by a bar) and *Relative-difference* (comparing the relative values of two entities represented by bars), whereas the categories for line graphs include *Big-jump* representing a sudden sharp change in value of an entity and *Change-trend-return* representing a trend that reverses and then returns to the original trend. The full set of categories can be found in [Elzer 2005] and [Wu et al. 2010].

A candidate intended message is constructed by instantiating the parameters of a message category with entities from the graphic. Conceptually, this is relatively straightforward for a bar chart. However, since line graphs do not consist of discrete entities but are instead a continuous set of connected points, we first abstract a line graph into a set of discrete entities (namely, visually distinguishable trends) from which the candidate messages can be constructed).

4.2.1. Abstracting a Line Graph into Visually Distinguishable Trends. Many line graphs, such as the one in Figure 2, consist of many short rises and falls, but a viewer summarizing it would be likely to regard it as consisting of a short overall stable trend from around 1900 to approximately 1930, followed by a long rising trend (both with high variance). As observed by Zacks and Tversky [Zacks and Tversky 1999], this tendency to associate lines with trends exists in part because of cognitive naturalness[6] and in part because of ease of perceptual processing. In fact, the cognitive "fit" of the line graph for representing trends is upheld by multiple findings from the basic Gestalt principles to the Wickens and Carswells' Proximity Compatibility Principle (grouping objects that are meant to be processed together) [Wickens and Carswell 1995] to Pinker's model of graph comprehension [Pinker 1990]. Thus, rather than working with the raw line graph, we developed a Graph Segmentation Module that divides a line graph into a sequence of visually distinguishable trends from which candidate intended messages can be constructed.

The Graph Segmentation Module applies a top-down recursive decision process. Starting with the initial line graph as a single segment, it applies a learned model to make a decision about whether the segment represents a single visually distinguishable trend or whether it should be split into two subsegments. In the latter case, the segment is split at the point that is furthest from a straight line connecting the end points of the segment and then the decision process is applied to each of the subsegments.

The model that makes the split/no-split decision on a segment was learned using a support vector machine [Tan et al. 2005] that considered 18 attributes [Wu et al. 2010]. Many of the attributes represent the result of statistical tests that estimate how well the segment represents a linear regression of the sampled points in the segment. The following are two such attributes:

[6] The term *cognitive naturalness* was coined by Zacks and Tversky [Zacks and Tversky 1999]; it means *requiring less mental work*.

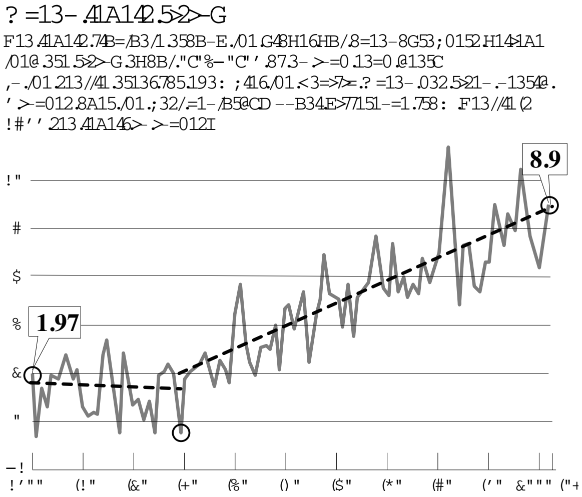


Fig. 11: Segmentation result for the graphic in Figure 2. Regression lines representing the visually-distinguishable trends are shown as dashed lines; the split points for segmenting the line graph are shown as circles.

— Attribute: Correlation Coefficient: This attribute is the absolute value of the Pearson product-moment correlation coefficient [Rodgers and Nicewander 1988]; it measures the tendency of the dependent variable to have a linearly rising or falling relationship with the independent variable. It is computed as the covariance of the X, Y values divided by the product of their standard deviation. The closer the absolute value of the correlation coefficient is to 1, the more the points represent a linear increasing or decreasing relationship. We hypothesized that the correlation coefficient might be helpful in determining whether a long set of short jagged segments, such as those between 1930 and the end of the graph in Figure 2, should be captured as a single rising trend and thus not be split further.

— Attributes: Runs Test: Five attributes are based on the Runs Test [Bradley et al. 1995]. A run is a sequence of consecutive sampled points that all fall above the regression line or all fall below the regression line. The number of runs is then compared with an estimate of the expected number of runs R_{mean} and its standard deviation R_{SD}; if the actual number of runs exceeds $(R_{mean} - R_{SD})$, then the Runs Test suggests that a least squares linear regression is a good fit for the data points and that the segment should not be split. We hypothesize that the Runs Test might be helpful when a segment consists of more than two trends. Five attributes are based on the Runs Test, including the result of the Runs Test (a binary value) and the actual number of runs R.

In contrast with other segmentation algorithms, we also include global attributes that consider the segment in the context of the overall line graph. The following is such an attribute:

— Attribute: Proportion: This attribute measures the fraction of the entire line graph represented by this segment. We hypothesize that segments that comprise a large portion of the graph are more likely to consist of visually distinguishable subsegments and thus should be split.

Evaluator	t value	significance level
Judge-1:	22.69	.001
Judge-2:	4.87	.001
Judge-3:	18.89	.001
Judge-4:	14.39	.001
Judge-5:	14.61	.001
Judge-6:	11.10	.001
Judge-7:	27.80	.001

Table I: Results of Welch's t-test

To evaluate our Graph Segmentation Module, we applied it to a corpus of 234 line graphs. Seven human evaluators were given the segmentations produced by the Graph Segmentation Module for each of the 234 line graphs, intermixed with bad segmentations for 20 additional graphs; the evaluators were not told that the set included intentionally bad segmentations. The evaluators were asked to rate each segmentation according to how well it captured the visually distinguishable trends in the graphic, using a scale from 1 to 5: 5 = ideal, 4 = very good, 3 = acceptable, 2 = poor, 1 = terrible. To assess the degree of agreement among the seven evaluators, we computed the Fleiss kappa statistic [Fleiss 1971] which is used in place of the Cohen kappa statistic [Cohen 1960] when the number of raters is greater than 2; the kappa statistic attempts to measure the degree of inter-rater agreement while discounting for the likelihood of chance agreement. The Fleiss' kappa statistic is 0.418 for the set of 254 line graphs which is regarded as moderate agreement. Another statistic that can be used for assessing degree of agreement among more than two raters is Kendall's coefficient of concordance W [Kendall and Babington-Smith 1939; Daniel 1989]; this statistic is more appropriate for rating tasks than the Kappa statistic since scores of 4 and 5 are regarded as much closer to agreement than scores of 4 and 1. The W value (taking into account ties in the ratings) for our 7 evaluators is .607 showing that there is a positive correlation of agreement among the raters.

The average rating for the segmentations produced by our system was 4.25 with a .55 standard deviation, showing that the performance of our model is between *very good* and *ideal*. The average rating for the bad segmentations that were added to the corpus was 1.57 with a standard deviation of .44, which was between *poor* and *terrible*. To determine whether the judge's ratings for the system's segmentations were significantly better than their ratings for the poor segmentations that were deliberately added to the evaluation set, we applied Welch's t-test which is appropriate for two samples having unequal variances. The results given in Table I indicate that the difference between the two sets of ratings were significantly different for all judges at the .001 level. These results show that our Graph Segmentation Module is successful at dividing a line graph into visually distinguishable trends.

As an example of the Graph Segmentation Module, consider the results shown in Figure 11 for the graph in Figure 2. The original line graph is split into two visually distinguishable trends, one from the beginning of the graph to around 1930 and the second from around 1930 to 2003. The split points are depicted by circles in Figure 11; the visually distinguishable trends are located between each pair of adjacent splitting points and are represented by dashed regression lines.

4.2.2. The Suggestion Module. A Suggestion Module is responsible for constructing a set of candidate intended messages. Each of the identified message categories has a set of parameters which can be instantiated with the label of a bar in a bar chart or the x-value of a point on a line graph to produce candidate intended messages.

Since the Graph Segmentation Module abstracts a line graph into a sequence of visually distinguishable trends, the Suggestion Module uses this sequence to construct a set containing every possible instantiation of the message categories for line graphs. For example, the category *Change-trend-return* captures messages that convey a change over a sequence of three trends, where the direction (rising, falling, or stable) of the second trend is different from that of the first trend and the third trend has the same direction as the first trend — i.e., the third trend resumes the direction of the first trend. Thus the *Change-trend-return* category takes four values as parameters: the x-value of the starting points of each of the three trends and the x-value of the ending point of the third trend. So for every sequence of three visually distinguishable trends produced by the Graph Segmentation module where the signs of the slopes of the first and third trends are the same and the sign of the slope of the middle trend is different, an instantiated candidate *Change-trend-return* message is produced.

Due to memory and processing limitations, it was not possible to consider all possible instantiations of the message categories for bar charts. For example, if a bar chart contained 10 bars, just considering every possible instantiation of the message category *Relative-difference* would produce 45 candidate messages, one for each pair of bars.[7] As noted in Section 4.1, the relative effort required for a perceptual task serves as a signal about whether the viewer was intended to perform that task, and the graph designer uses attention-getting devices to make certain entities salient in the graph. Thus we use relative perceptual effort and salience as heuristics to limit the number of candidate intended messages for bar charts. The 10 perceptual tasks that represent the least effort for the bar chart are identified, along with every perceptual task involving a salient entity in the bar chart; each instantiation of a message category that requires performing one of these perceptual tasks is included as a candidate message.

4.3. Reasoning with a Bayesian Network

A Bayesian network is used to determine the intended message of a graphic. The conditional probability tables for the Bayesian network are learned from a training corpus of graphics. Given a new graphic, a Bayesian network is dynamically built as described below, based on the candidate messages that were discussed in Section 4.2. Once the evidence from the extracted communicative signals has been entered into the network and the probabilities propagated through the network, the candidate message with the largest probability is selected as the system's hypothesis about the graphic's intended message.

The structure of the network facilitates the necessary reasoning. The top level node of the network represents the set of message categories; the values in this node are the twelve message categories, as shown in Figure 12 for simple bar charts. Each of these categories then appears as a child of the top-level node. [8] The children of each message category node are the set of instantiated candidate messages in that category. For bar charts, the network is further expanded to include its requisite perceptual tasks as children.

[7] As discussed in Section 4.3, we use a Bayesian network to arbitrate among the candidate intended messages. Memory problems are caused not only by the number of nodes in the network but also the size of the conditional probability tables resulting from inhibitory links between exclusive alternatives in the network.

[8] This additional level does not affect the inference process or the computation of the probabilities in the network. Its presence merely simplifies the conditional probability tables in the network by limiting the size of the tables that would typically occur just below the top-level node; those tables now only need to contain the InPlan and NotInPlan probabilities of their parent node, rather than the InPlan and NotInPlan probabilities for all the values of the top-level node.

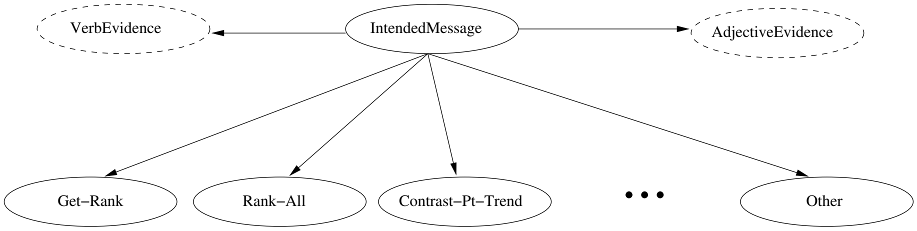


Fig. 12: Top Level of the Bayesian Network for Bar Charts

A Bayesian network requires evidence in order to hypothesize the intended message for the graphic. Perceptual task effort is a clue about whether a specific perceptual task was intended by the graph designer; thus relative perceptual task effort (low, medium, high) as computed by the rules described in Section 4.1 is inserted as evidence beneath each perceptual task node. Salience is a clue about whether a specific entity was intended to play a significant role in the intended message — that is, whether the entity is a parameter in the intended message. Thus the presence or absence of each kind of attention-getting device for the parameters (coloring, annotation, mention in the caption, etc.) is attached as evidence beneath an instantiated candidate message. On the other hand, whether or not a word from one of the verb categories appears in the caption serves as evidence about the general category of intended message and thus appears as a child of the top-level node in the network.

Figure 12 shows the top level of the Bayesian network for bar charts, with the verb and adjective evidence nodes attached to the top-level IntendedMessage node. Beneath this are nodes for each of the possible message categories. The rest of the Bayesian network is built dynamically for each new graphic. For example, if there were three instantiated candidate messages for the Get-Rank message category, then the Get-Rank message category node would have three children as shown in Figure 13; the dashed lines are inhibitory links enforcing mutual exclusivity among the three instantiated candidate messages. Figure 14 displays the subnetwork for one of these candidate messages, namely Get-Rank(BAR1). It captures the fact that one could get the rank of a bar in one of two ways: 1) by perceptually noticing the bar and then perceiving its rank or 2) by finding the label of the desired bar and then perceiving that bar's rank. Their children capture tasks that must be performed to get a bar's rank in one of these two ways. The children of the leaf nodes in Figure 14 are evidence nodes capturing

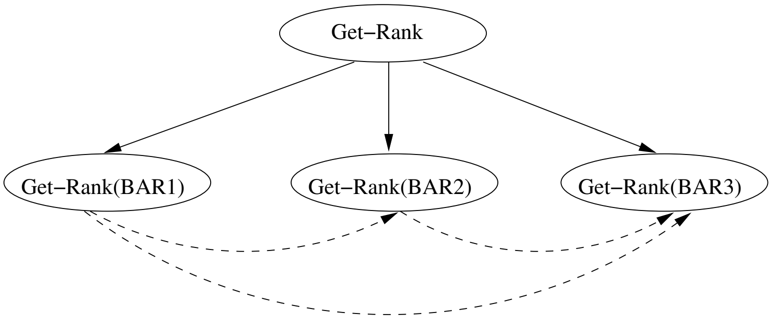


Fig. 13: Possible Instantiations of Get-Rank

the communicative signals that serve as evidence for whether the primitive perceptual task represented in the leaf node is part of recognizing the graphic's intended message. Figure 15 displays the evidence nodes for the Perceive-Rank(BAR1) primitive perceptual task. Each node in the Bayesian network has a conditional probability table capturing the probability of each of the node's values given the values of its parent nodes; these are the probability tables that were learned from the training corpus. Further explanation of the structure of the Bayesian networks for bar charts and for line graphs is given in [Elzer et al. 2005] and [Wu 2012] respectively.

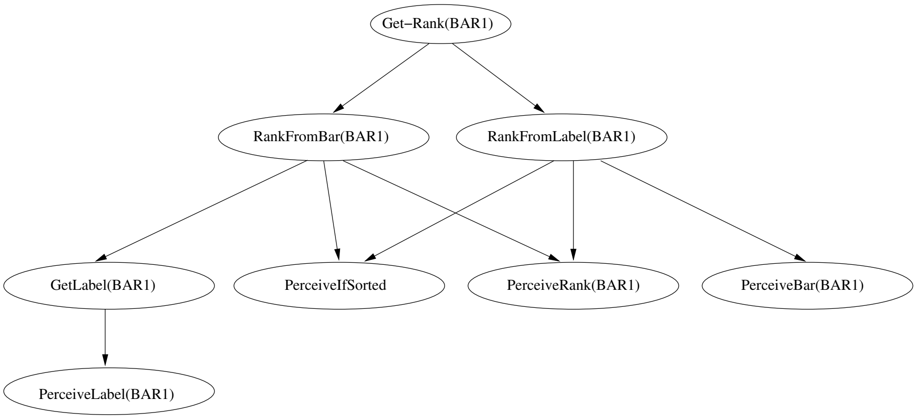


Fig. 14: Get-Rank(BAR1) Subnetwork

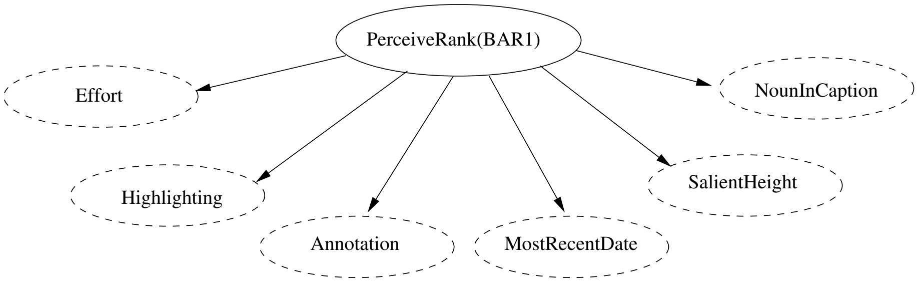


Fig. 15: Perceptual Task Node with Evidence Nodes

4.4. Evaluation and Examples

Our system for recognizing the intended message of an information graphic was tested on a corpus of 110 bar charts and 240 line graphs collected from various popular media such as *Newsweek*, *Business Week*, *USA Today*, and local newspapers. Each of the graphics was annotated with its intended message by two human coders using consensus-based annotation [Ang et al. 2002], in which the coders independently identified the intended message for each graphic but then discussed graphics where their annotations differed in order to come to agreement as to its intention.

Using leave-one-out cross validation in which one graph served as the test graph and the other bar charts or line graphs were used to compute the conditional probability

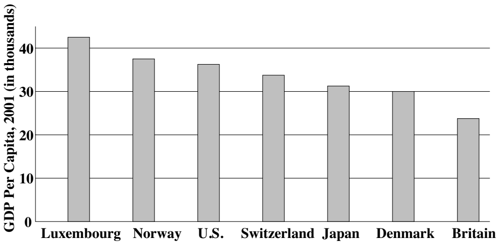


Fig. 16: A bar chart with no salient entities.

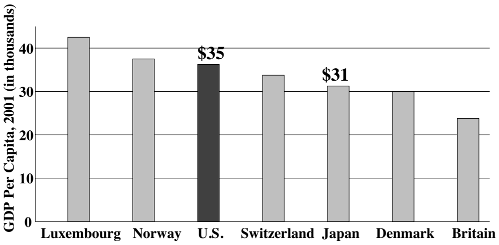


Fig. 17: A bar chart with two salient entities.

tables for the network, our system's accuracy in recognizing the correct intended message (both message category and instantiated parameters) for bar charts was 79.1% and for line graphs was 72.08%. The 95% confidence interval [Witten and Eibe Frank 2011] for bar charts is (70.6%,85.7%) and for line graphs is (66.1%,77.4%). The lower success rate for line graphs is partially due to the fact that errors from the Graph Segmentation Module propagate into the parameters of the candidate messages produced by the Suggestion Module and thus the Bayesian network. For example, on our test set of 240 line graphs, the correct segmentations are produced for only 215 of them. Thus a soft upper bound[9] on the success rate of the overall message recognition system is 89.5% for the test set. For the purposes of the SIGHT system, it is important to note that even if the message recognized by our Bayesian network does not match the intended message identified by the human annotators, the message recognized by our system still contains information conveyed by the graphic. For example, if the system hypothesizes a *Change-trend* message when the intended message is really a *Change-trend-return*, the data captured by the graphic will include a changing trend; but the system has failed to correctly identify the overall message that the graphic is intended to convey.

As examples of the messages recognized by our system, consider the bar chart in Figure 1 and the line graph in Figure 2. Our system correctly recognizes the intended messages of these graphics, namely that American Express ranks third in terms of total credit card purchases each year (the logical representation is Rank(American Express, 3)) and that there is a changing trend in annual difference from Seattle's 1899 sea level (the logical representation is Change-trend(1901, stable, 1928, rise, 2003))[10] respectively.

To see the impact of varying evidence in the graphic, consider the two graphs in Figures 16 and 17. Our system hypothesizes a Rank-all intended message for the graphic in Figure 16 — namely, that it is intended to convey the relative rank of the various countries depicted in the graphic. However, the bars for the US and Japan are both salient in the graphic in Figure 17, with the bar for the US more salient since it is both colored differently and annotated with its value. These communicative signals suggest that both bars play a prominent role as parameters in the graphic's intended message, and our system hypothesizes the intended message to be the relative difference between the GDP of the US and Japan, whose logical representation is Relative-difference(US,Japan).

[9] If the Graph Segmentation Module produces an incorrect segmentation, it is possible that the correct intended messsage could still be identified if the segments denoting the parameters of that message were correctly represented in the incorrect overall segmentation.

[10] Interpolation is used to produce the exact x-values for the points on the graph identified as split points.

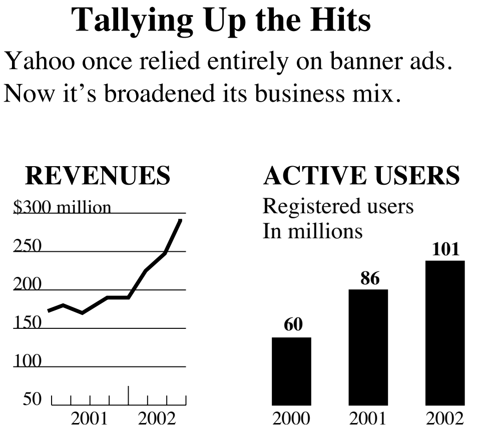


Fig. 18: A Composite Graphic from Newsweek.[11]

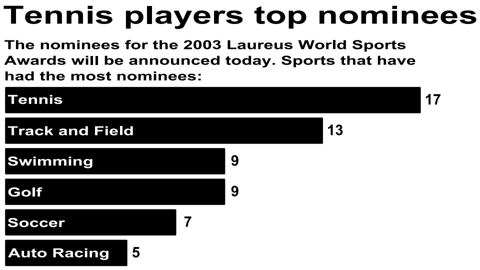


Fig. 19: A Graphic from USA Today.

4.5. System Limitations

The most significant factors that impact the system's success in recognizing a graphic's intended message are the quality of the representation produced by the Visual Extraction Module and the accuracy of the segmentation of a line graph into visually distinguishable trends. Currently, SIGHT works with electronic documents whose graphics are clearly drawn and represented as .gif files, have standard representations of graphical elements, and are in a fixed set of fonts. Thus graphics that have non-standard representations (such as using icons for bars) cannot be handled. In addition, as noted in Section 4.2.2, memory limitations prevent the system from considering all possible instantiations of the message categories for bar charts. Thus occasionally the correct intended message for a bar chart might not be among the candidate messages constructed by the Suggestion Module and consequently the system cannot recognize it as the graphic's intended message. The successful extraction of evidence also affects system performance, but to a lesser extent. For example, we use a part-of-speech tagger and parser to identify helpful verbs in the caption and nouns that match labels in the graphic. If a helpful noun or verb is not extracted, less evidence is available for identifying the intended message. This is exacerbated when labels do not exactly match nouns in the caption, such as a label that says USA and a caption that says United States. These are all issues that must be addressed in future work.

5. IDENTIFYING THE MEASUREMENT AXIS DESCRIPTOR

The above logical representations of the intended message do not yet capture what is being measured in the graph, such as *total credit card purchases each year* for the bar chart in Figure 1. Unfortunately, many information graphics do not explicitly label the dependent axis with what is being measured in the graph. For example, the line graph in Figure 2 shows a changing trend in *annual difference from Seattle's 1899 sea level*, the bar chart in Figure 18 shows a rising trend in *Yahoo's registered users*, and the bar chart in Figure 19 conveys that tennis ranks first in terms of *number of nominees for the 2003 Laureus World Sports Awards*; yet none of these information graphics has its y-axis labeled with this referent. Thus it was necessary to identify a referent for what is being measured on the y-axis — what we refer to as the *measurement axis descriptor*.

We analyzed a large corpus of information graphics whose measurement axis descriptor was identified by two human annotators. We observed that

- the measurement axis descriptor consists of a core noun phrase or wh-phrase appearing in the graphic's text, possibly augmented with words appearing elsewhere. For example, for the bar chart in Figure 18, *registered users* is the core of the measurement axis descriptor *Yahoo's registered users*.
- the textual elements of the graphic form a hierarchy based on placement in the graphic, and the core of the measurement axis descriptor typically appears in the lowest text element present in the graphic that is not solely a unit or scale indicator. Moreover, if the text element contains multiple sentences or sentence fragments, the core of the measurement axis descriptor typically appears near the end of the textual element.
- certain cues can be identified that indicate the core of the measurement axis descriptor. For example, if a sentence at a text level begins with a "Here is" or "Here are" cue phrase, the core typically follows as a noun phrase or the object of a prepositional phrase. Similarly, if a fragment at a text level consists solely of a noun phrase that is not a proper noun, that noun phrase is typically the core of the measurement axis descriptor. This latter observation accounts for the core of the measurement axis descriptor for the bar chart in Figure 18.

Motivated by this corpus analysis, we developed a set of heuristics[12] that construct a measurement axis descriptor by melding together words and phrases extracted from the text that appears in the graphic. We organized the textual elements of the graphic into a hierarchy from lowest to highest:

(1) Dependent_Axis_Label: any text explicitly labeling the dependent axis
(2) Text_In_Graphic: any text appearing within the graphic itself, such as the phrase *Total Credit Card Purchases Per Year* that appears within the graph in Figure 1.
(3) Description: any text appearing beneath the main caption of the graphic, such as the several sentences underneath the caption *Ocean levels rising* in Figure 2.
(4) Caption: the text comprising the lead caption on the graphic, such as *Ocean levels rising* in Figure 2.
(5) Overall_Description: the text appearing beneath the main caption of a composite graph that contains several individual graphs, such as the two sentences beneath the caption *Tallying Up the Hits* on the composite graph in Figure 18.
(6) Overall_Caption: the text comprising the lead caption on a composite graph, such as *Tallying Up the Hits* in Figure 18.

We constructed an ordered set of heuristics that were applied first to the lowest level text element in the hierarchy and then in turn to higher elements of the hierarchy until a core was identified.[13] The following are three sample heuristics:

- If the current sentence at the text level begins with a "Here is" or "Here are" cue phrase, the core is the object of the prepositional phrase (if any) following the cue phrase; otherwise, the core is the noun phrase following the cue phrase.
- If a fragment at a text level consists solely of a noun phrase that is not a proper noun, that noun phrase is the core.

[11] This figure displays two of the five individual graphs comprising the composite graphic that appeared in *Newsweek*.

[12] Our corpus analysis was limited to graphics from articles written in American English and thus our heuristics are not necessarily applicable to graphics from articles in other languages.

[13] If a text level contained multiple sentences and sentence fragments, the ones at the end were examined first.

— If the text element is a sentence, the noun phrase preceding the verb phrase in the sentence is the core.

Postprocessing then revises or augments the core to get the complete measurement axis descriptor. If the core has a modifier (such as *rising* or *changing*) that matches an intended message category, that adjective is removed since it relates to the intended message, not what is being measured on the dependent axis. For example, the core for the graphic in Figure 10 is initially *"Declining Durango sales"* which is reduced to *"Durango sales"*. If the extracted core was a complex noun phrase whose head matches the ontological category of the labels, the head should not be part of the core since it describes the independent axis, not the dependent axis. This is the case for *sports that have had the most nominees* which would be the extracted core for the bar chart in Figure 19. Thus in such cases, the nouns and subsequent prepositional phrases in the modifier are collected as the core; for the bar chart in Figure 19, this would produce *nominees* as the new core. Several augmentation rules are then applied to augment the core with other text extracted from the graphic. One such augmentation rule specializes a noun phrase as follows:

— If there is only one proper noun at the same level and all text levels higher in the hierarchy than the text level from which the core was extracted or the Caption or Overall_Caption contains only one proper noun, then the possessive form of that proper noun is appended to the front of the core as long as the proper noun is not a label on the independent axis of the graph. Thus for the graphics in Figures 18 and 19, the measurement axis descriptors become respectively *"Yahoo's registered users"* and *"Laureus World Sports's nominees"*.

Finally the unit of measurement is added if it is not already part of the identified descriptor. Thus the final measurement axis descriptors for the graphs in Figures 18, 19, and 2 are *"the number of Yahoo's registered users"*, *"the number of Laureus World Sports's nominees"*, and *"annual difference from seattle's 1899 sea level, in inches"*. The full set of heuristics and postprocessing rules can be found in [Demir 2010].

An evaluation of the methodology was conducted on a new corpus of 205 randomly selected bar charts from 21 different newspapers and magazines.[14] Two human evaluators rated the identified measurement axis descriptor on a scale from 5 to 1: 5 (excellent text), 4 (very good, interpreted as understandable but awkward), 3 (good, contains the right information but is hard to understand), 2 (poor, missing important information), 1 (very bad). They also evaluated three baselines which used respectively the text appearing as the Dependent_Axis_Label, the Text_In_Graphic, and the Caption. The Kappa statistic for inter-rater agreement was .79 for the measurement axis descriptors produced by the system, which represents substantial agreement. The Kappa statistics for the baselines were respectively .89, .92, and .91. These Kappa statistics appear unusually high. However, the measurement axis descriptors produced by the system were generally either perfect or very bad, making the ratings of the evaluators the same. Agreement was even greater for the baselines; this is accounted for by the fact that the descriptors provided by the baselines were overwhelmingly terrible or non-existant (both leading to a score of 1). For example, consider the graphic in Figure 19. The graphic does not have a dependent axis label or any text in graphic; thus if it were part of the test set, it would have received a score of 1 for both of the first two baselines. Similarly, for the baseline that uses the caption, it would have received a score of 1 from both evaluators since *"Tennis players top nominees"* is nowhere

[14] Although we use the same methodology for identifying the measurement axis descriptor for line graphs, we have not yet evaluated it for line graphs.

near the desired measurement axis descriptor of *"the number of Laureus World Sports nominees"*.

If the evaluators differed in their ratings for a graphic, the lower rating was recorded for the measurement axis descriptors produced by the system and the higher rating was recorded for the baselines. The evaluation score for the measurement axis descriptors produced by the system was 3.574 which is midway between *good* and *very good*. This score was far better than the scores (1.475, 1.757, and 1.876) for the three baselines and a Welch's t-test shows that the difference between the system's ratings and the ratings for the three baselines is significant at the .001 level.

6. COHERENT INCLUSION OF INFORMATION GRAPHICS

Our system uses JAWS screen-reading software to read the text of the article to a user and prompt the user about the presence of an information graphic; it then accepts requests for access to the graphic's content. Unfortunately, in popular media, information graphics often don't appear adjacent to paragraphs to which they are related. And unlike scientific articles where the information graphic will be referred to explicitly, such as "As shown in Figure 1," articles in popular media generally lack explicit references to their information graphics. These two characteristics of popular media mean that geographical location in the article and explicit references to a graphic cannot be relied upon to determine where the content of an information graphic most coherently fits into the article text. While determining the best place in the reading of a document to render an information graphic requires input from people with visual impairments, this study cannot be carried out unless the system can identify the paragraph most relevant to the graphic (since presumably one might want to render the graphic either before or after this most relevant paragraph and it thus must be one of the available options in the study). Thus, the SIGHT system took on the problem of identifying the portion of text that is most relevant to an information graphic.

6.1. Methodology:P-KL and P-KLA

We developed a basic method P-KL for identifying the paragraph most relevant to an information graphic, based on Kullback-Leibler divergence (referred to as KL-divergence), and then augmented it to produce a more sophisticated method P-KLA. KL-divergence is widely used in statistics and natural language processing to measure the asymmetric distance of two probabilistic distributions or models. It can be expressed as:

$$D_{KL}(p||q) = \sum_{i \in V} p(i) log \frac{p(i)}{q(i)} \tag{1}$$

where i is the index of a word in vocabulary V, and $p(i)$ and $q(i)$ are the probabilities of word i in two distributions of words. Intuitively, document (paragraph) p is less divergent from document q if the words in p appear in about the same relative frequency as they do in q. Liu et al. [Liu and Croft 2002] applied KL divergence on passages to improve the accuracy of document retrieval. For our task, p is a smoothed word distribution[15] built from a combination of three of the graph's textual components (see Section 5): Caption which is the main title for the information graphic, Description which is any additional text that elaborates on the caption (such as the sentences in Figure 2 that lie immediately beneath the caption *Ocean levels rising*), and Text_In_Graphic which is any text appearing inside the graphic area. q is another

[15] Smoothing substitutes small pseudo-counts for words with no occurrences in order to address a sparse data problem.

smoothed word distribution built from a paragraph in the multimodal document. We rank the paragraphs by their KL divergence scores from lowest to highest, since lower scores indicate a higher similarity.

The basic method only considers the textual components associated with the information graphic. But an information graphic consists of two parts: the textual part and the graphic part. The textual part of the graphic generally has a lot of words having to do with the domain of the graphic. In contrast, one can talk about the graphic part of an information graphic in a very domain-independent way. For example, the graphic may present trends (rises or falls), results (higher or lower), or (in the case of bar charts) ranks or comparisons. We would like our system to be able to recognize words such as these as being about a graphic. Thus we decided to explore whether we could automatically extract a set of expansion words that are commonly used in paragraphs that are relevant to information graphics and thus serve as signals that the paragraph is related to a graphic. Method P-KLA would then augment the textual component of a graphic with these expansion words before applying KL-divergence to identify the most relevant paragraph.

To construct this word set, we apply an iterative process on a training set of 395 information graphics with full articles from multiple national sources such as *USA Today*, *Businessweek*, *News Week*, *New York Times*, and *Wall Street Journal* and some local sources such as *The Wilmington News Journal*. First, for each information graphic in our training set, we use KL divergence to identify m pseudo-relevant paragraphs in the document. This is similar to the pseudo relevance feedback technology used in information retrieval [Zhai 2008], except that the information retrieval process considers a single query whereas we are using a set of information graphics and associated documents to identify an expansion set that can be applied to all information graphics. If there are N information graphics, we produce a set of at most $m \cdot N$ relevant paragraphs. In our experiment, we arbitrarily chose $m = 3$.

The next step is to extract a common word set from the set of pseudo-relevant paragraphs. We assume that the collection of pseudo relevant paragraphs was generated by two models, one producing words relevant to the information graphics and one producing words relevant to the topics of the documents. Let W_g represent the word frequency vector that generates words relevant to the information graphics, W_a represent the word frequency vector that generates words relevant to the domains of the articles, and W_p represent the word frequency vector of the pseudo-relevant paragraphs. We can compute W_p from the pseudo-relevant paragraphs, and we can estimate W_a as the word frequency vector for the entire set of articles. We want to compute W_g by filtering the components of W_a from W_p. The problem can be formulated making the following assumptions:

(1) $W_p = \alpha W_a + \beta W_g$ where $\alpha > 0$ and $\beta > 0$, which means the word frequency vector for the pseudo-relevant paragraphs is a linear combination of the background (topics) word frequency vector and the graphic word vector.

(2) $< W_a, W_g >= 0$ which means the background word vector is orthogonal to the graph description word vector. Here we assume that the vast majority of the article concerns the domain of the article and the graphics words are primarily restricted to discussion of the graphics - a very small portion of the article. Thus W_a will be completely dominated by the domain information and thus we can consider that the graphic word vector is independent of the background word vector and these two share minimal information. Since we use a vector space model to represent W_a and W_g, orthogonality is obtained by assuming that these two word vectors have minimum similarity.

(3) W_g is assumed to be a unit vector. Whether or not W_g is a unit vector is immaterial for our method, since we are interested only in the relative rank of the word frequencies, not their actual values. However, assuming that W_g is a unit vector gives us three equations in three unknowns (W_g, α, and β) which can be solved for W_g.

With these three assumptions, we obtain the following results:

$$\alpha = \frac{< W_p, W_a >}{< W_a, W_a >} \quad (2)$$

$$W_g = \text{normalized} \left(W_p - \frac{< W_p, W_a >}{< W_a, W_a >} W_a \right) \quad (3)$$

After we compute W_g, we use WordNet to filter out words whose dominant sense is neither *verb* nor *adjective*, under the assumption that nouns will be relevant to the domains or topics of the graphs (and are thus *noise*) whereas we want a general set of words (such as *"increasing"*) that are typically used when writing about the data in graphs. To roughly estimate whether a word is predominantly a verb or adjective, we determine whether there are more verb and adjective senses of the word in Wordnet than there are senses that are nouns. We rank the words in the filtered W_g by their frequency and select the k (we chose $k = 25$ in our experiments) most frequent words as our candidate expansion word list.

Since the textual components were used to identify pseudo-relevant paragraphs and then pseudo-relevant paragraphs (as opposed to truly relevant paragraphs) were used to construct the candidate word list for expanding the textual components, the accuracy of both the pseudo-relevant paragraphs and the candidate expansion word list are suspect. Thus we apply the two steps (identify pseudo-relevant paragraphs and then extract a word list for expanding the textual components) iteratively until convergence or minimal changes between iterations.

Thus far, we have the textual component of the graphic and a set of domain independent expansion words to use in identifying paragraphs that are relevant to an information graphic. One piece of textual information not taken from the graphic is the labels in the axes that correspond to points in the graphic because these are generally too numerous or cover too wide a span as to be targeted specifically to a discussion of the graphic. However, some of these points may be stressed by the graphic and thus be very important to its description. These are points relevant to the intended message. Therefore, we also augment the textual component with the parameters of the graphic's intended message that match points on the x-axis. For example, the logical representation of the intended message recognized by our system for the line graph in Figure 2 is Change-trend(1901, stable, 1928, rise, 2003) which means that the line graph conveys a changing trend in ocean levels over the period from 1901 to 2003 with the change from relatively stable to rising occurring around 1928; the parameters 1901, 1928, and 2003 are added to the textual component.

Thus our second method, P-KLE, applies KL-divergence to the textual component, augmented with both the expansion word list and the parameters of the intended message, to rank relevant paragraphs from an article. Because the textual component may be even shorter than the expansion word list, we don't add a word from the expansion word list to the textual component unless the compared paragraph also contains this word; otherwise the impact of the words originally in the textual component will be severely diluted.

Using the most relevant paragraph in presentation must be evaluated with people with visual impairments. However, in the current implementation, when a line graph is encountered, SIGHT uses method P-KLE to identify the paragraph in an article that is most relevant to an information graphic and prompts the user about the presence

	geographically closest	P-KL	significance level over geographically closest	P-KLE	significance level over P-KL
TOP	0.25	0.45	0.001	0.57	0.01
COVERED	0.37	0.62	0.001	0.67	not significant

Table II: Success rates for baseline method and methods P-KL and P-KLE

of that graphic at the end of reading the paragraph. Thus SIGHT exhibits intelligent behavior by not merely providing access to a graphic at the point it appears in the document but instead attempts to integrate the graphic coherently into the article. SIGHT is being extended so that identification of the most relevant paragraph is also done for bar charts. The primary task is the construction of the set of expansion words that commonly appear in paragraphs relevant to bar charts which are likely different from words relevant to line graphs. For example, we anticipate that the set will include words such as *rank* since bar charts often convey the rank of several entities with respect to a particular criteria (such as the rank of airlines with respect to revenue).

6.2. Evaluation

To evaluate our methods for identifying paragraphs relevant to an information graphic, we used a test set of 100 line graphs and articles (all different from the training set used to construct the expansion word list). Two human evaluators identified paragraphs in each document that were relevant to its constituent information graphic and ranked them in terms of relevance. On average, Evaluator-1 selected 1.99 paragraphs and Evaluator-2 selected 1.67 paragraphs. For 66% of the graphs, the two evaluators agreed on the top ranked paragraph with a Kappa statistic of .624[16] this shows that in many cases, the most relevant paragraph is not obvious and that several possibilities exist.

The point of the method is to identify the most relevant paragraph as a logical place to insert a description of the graphic. Therefore, though our algorithm provides a ranked list of paragraphs in terms of relevance, only the top ranked one will be used. This leads to two evaluation criteria:

(1) TOP: the method's success rate in selecting *the most relevant paragraph*, measured as how often the most relevant paragraph identified by the method matches one of the two evaluator's top-ranked paragraph.
(2) COVERED: the method's success rate in selecting *a relevant paragraph*, measured as how often the most relevant paragraph identified by the method matches one of the paragraphs identified as relevant by the evaluators.

Thus TOP and COVERED evaluate respectively whether our method will present the graphic's content at **the most relevant point** in an article and **a relevant point** in an article.

We used the paragraph that was geographically closest to the information graphic as a baseline method for comparison with the paragraph selected by our system. Table II presents the results of the evaluation. P-KL and P-KLE select the best paragraph 45% and 57% of the time respectively, and select a relevant paragraph 62% and 67% of the

[16] Since the selection of paragraphs is different for each document, the probability of chance agreement in the Kappa statistic is computed assuming that the probability of selection is 1/n where n is the number of paragraphs in the document, as suggested by [Brennan and Prediger 1981] when either or both of the marginals can vary.

time respectively. Both of our methods outperform the baseline method. A *binomial test* shows that the improvements of P-KL over the baseline method and P-KLE over P-KL for selecting the most relevant paragraph are both statistically significant at the 0.01 significance level or better.

7. CONSTRUCTING THE INITIAL SUMMARY

Given the logical representation of a graphic's intended message and the referent for its measurement axis label, SIGHT can use FUF/SURGE [Elhadad and Robin 1999] to realize an English sentence conveying the graph's primary message. For example, for the line graph in Figure 2, SIGHT can produce the sentence *"This graphic conveys a changing trend in annual difference from seattle's 1899 sea level in inches, relatively stable from 1901 to 1928 and then rising to 2003"* which JAWS then reads to the user as synthesized speech. Although a graphic's intended message captures its primary high-level communicative goal, we hypothesize that the graphic may contain other salient propositions that elaborate on the intended message and are important to include in an initial summary. For example, the line graph in Figure 2 is very volatile with many sharp changes, and this feature presumably would be important to include in the initial summary. This hypothesis is supported by an informal experiment in which participants were asked to produce summaries for a series of line graphs [McCoy et al. 2001]; we observed that a description of the intended message was almost always included in graph summaries written by human subjects and that the subjects consistently augmented the intended message with descriptions of various salient visual features of the graphic (e.g., steepness of a trend line, volatility of data values). Although summaries of different graphics belonging to the same intended message category but having different visual features varied in the additional details provided by the human subjects, which visual features were found to be salient also depended on the graphic's intended message. The fact that these summaries did not include *all* information that could possibly be extracted from the graphic (e.g., the value of every point on the line graph), but rather only propositions corresponding to the visual features the subjects deemed most salient is consistent with Grice's Maxim of Quantity [Grice 1975]: utterances should be as informative as necessary, but not more so.

We conducted a series of data collection studies designed to determine which propositions human subjects expect to find in a summary of a graph.[17] In our initial work involving simple bar charts, we asked subjects to fill-out questionnaires in which they were tasked with assigning an importance rating to various propositions that could be extracted from a given bar chart. An analysis of the participants' responses led us to formulate a set of rules to determine the salience of various visual features; the rules are triggered in SIGHT based on thresholds established for each visual feature. However, due to the continuous nature of the data series in line graphs (as opposed to the discrete nature of bar charts), it was not possible to present subjects with an unbiased finite set of propositions to rate. Thus we performed a study in which human subjects were presented with a variety of different line graphs along with their intended message and asked to complete a brief summary augmenting the intended message with the important information they conveyed. This collection of summaries was subsequently examined by a human annotator who identified which propositions were included by the summary writers, and we formed rules for identifying the salience of informational propositions in line graphs based on these annotations. Again, due to the continuous nature of line graphs versus the discrete nature of bar charts, these new rules included a weighted importance score linked to the magnitude of the asso-

[17] This work considered the graph in isolation. As noted in Section 11, future work will consider the accompanying text and its impact on the salience of graphic features.

ciated visual feature (e.g., the degree of steepness or volatility), rather than operating according to thresholds as with bar charts.

7.1. Content Identification Rules for Bar Charts

To identify propositions (in addition to the intended message) that should be included in an initial summary of a bar chart, we constructed a set of 21 representative bar charts representing different categories of intended messages and visual features. We then conducted an experiment in which 20 human subjects were presented with subsets of these bar charts, the intended message of the bar chart, and a list of additional propositions that could be extracted from the bar chart (such as the value of the smallest bar). The subjects were asked to classify these propositions into one of three groups according to how important they thought it would be to include that proposition in a brief textual summary of the graphic. These three categories were: Essential (the proposition *should* be included in the summary), Possible (*could* be included, but not essential), and Not Important (*should not* be included). Since different participants assigned different levels of importance, we needed to identify the participants' general tendency. We assigned a numerical score to each importance level: Essential = 3, Possible = 1, and Not Important = 0.[18] The importance level of a proposition was computed by summing the scores of the importance levels assigned by the participants. For example, if three subjects rated a proposition as Essential and two as Possible, the importance score for that proposition was $(3 * 3) + (2 * 1) = 11$. We defined **majority importance level** as the importance level that would be obtained if half of the participants were to classify a proposition as essential. An individual proposition was labeled as **highly-rated** if its actual importance level was greater than or equal to the majority importance level.

The rules for selecting propositions to include in the initial summaries of bar charts were based on patterns observed between the intended message category, visual features of the graphic, and the highly-rated propositions identified by human judges. Two types of rules were constructed: Message-category-only rules and Message-category-visual-feature rules. If a proposition was marked as highly-rated for all graphs in a particular message category, then the choice to include that proposition depended only on the message category and not on any visual features:

- Message-Category-only:
  IF $Message - category = m$ and $Rating(P, m) = highly - rated$
  THEN select P

For propositions that were highly-rated for only a particular subset of the graphics in a given message category, we identified a visual feature present in graphics where the proposition was highly-rated and missing in graphics where it was not rated highly. A second set of content identification rules used the presence or absence of this feature to determine when to include the proposition.

- Message-Category-Visual-Feature:
  IF $Message - category = m$ and $Visual - feature = v$
  and $Rating(P, m, v) = highly - rated$ and $Rating(P, m, \neg v) = \neg highly - rated$
  THEN select P

For example, in content identification rule #1 below, the message category of the graphic is the only determining factor in deciding whether or not to include the proposition in the initial summary. However, in rules #2 and #3, both the message category

[18] The reason for the (3,1,0) scale is to guarantee that a proposition will not be labeled as a highly rated proposition if none of the participants classified it as essential.

and the presence of a visual feature (e.g., "not steady," "within range") are used in the determination process.

(1) If (message_category **equals** 'increasing trend') then **include**(proposition conveying the rate of increase of the trend):
Include the proposition conveying the rate of increase of the trend for all bar charts having an 'increasing trend' intended message.

(2) If [(message_category **equals** 'increasing trend') and **not_steady**(trend)] then **include**(proposition conveying the period(s) showing a decrease):
Include the proposition conveying the period(s) showing a decrease (i.e., the exceptions) for all bar charts having an 'increasing trend' intended message but which are not steady.

(3) If [(message_category **equals** 'rank all') and (**all**(**value**(bar)) **within_range**((0.7 * **average**(all bars)), (1.3 * **average**(all bars))))] then **include**(proposition indicating that the bars vary only slightly):
If the values of all bars are close in a bar chart having a 'rank all' intended message, then include the proposition indicating that bar values vary only slightly.

For rules involving visual features (e.g., #2 and #3 above), discrete thresholds needed to be defined in order to determine whether the feature was present in a given bar chart. For rule #2, a trend was considered "not steady" when there was at least one bar with a value lower than the previous bar within an otherwise overall increasing trend. Rules such as #2 are consistent with the Joshi/Webber/Weischdel maxim that responses should be modified to avoid the hearer drawing false conclusions [Joshi et al. 1986] — in this case, the false conclusion that the trend is steady. For rule #3, the values of all bars had to be within 70-130% of all other bars in order to select the proposition indicating that bar values vary only slightly. A total of 31 content identification rules were constructed (some covering multiple message categories), enabling propositions to be selected for inclusion in the initial summary.

For example, given the bar chart in Figure 20, the following propositions are selected for inclusion in the initial summary (in addition to a description of the overall intended message):

— The rate of increase of the trend, which is slight[19]
— The fact that the trend is not steady and has variability
— The overall percentage increase in the trend, which is 225%

7.2. Content Identification Rules for Line Graphs

To identify propositions (in addition to the intended message) that should be included in an initial summary of a line graph, we selected 23 different line graphs covering the eight most common intended message categories and different visual features. Ten of the graphs were real world examples taken directly from popular media sources. Another ten graphs were adapted from real world graphs – modified to isolate visual features in order to study their individual effects. The final three graphs were specially-created to fill a gap in the coverage of message categories and visual features for which no good example was available (e.g., Figure 21). 69 human subjects were given all 23 line graphs (in different orders), each of which was accompanied by an initial sentence conveying its intended message as identified by a human annotator. For example, the initial sentence given for Figure 21 conveyed the changing trend in Boscov's jacket sales — namely, that sales rose from November to February and then fell through May. Participants were tasked with writing additional sentences so

[19] This refers to the fact that the year-by-year change is small

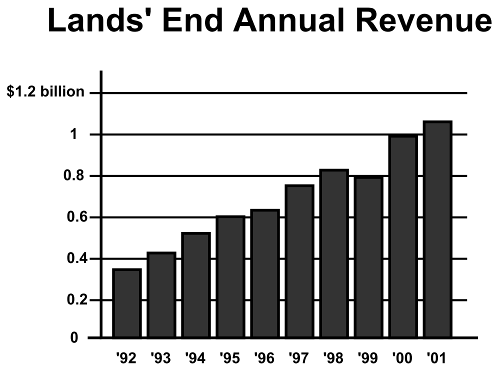


Fig. 20: Sample bar chart processed by SIGHT.

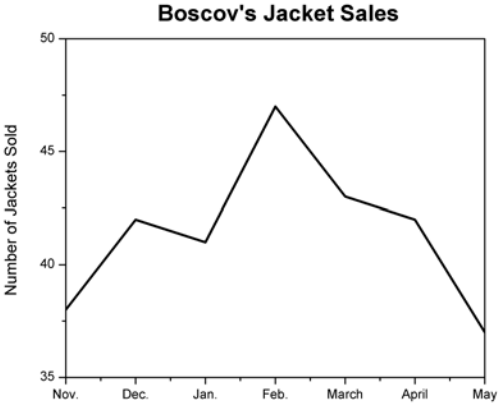


Fig. 21: Sample line graph created for this study.

that the completed summary would be an appropriate summary for someone who was blind that came across the graphic while reading popular media. The summary should capture the most important information conveyed by the line graph, and should not be misleading. They were also aware that the initial summary that they wrote need not be exhaustive because a blind reader would have the opportunity to ask follow-up questions if they desired. The participants wrote summaries for as many or as few of the full set of 23 line graphs as they wished during a single one-hour session.

In total, 965 summaries were collected, with between 37 and 49 summaries for each individual line graph. There was an average of 2.26 sentences per summary, in addition to the initial sentence describing the intended message that was given to the participants. Two example summaries collected for Figure 21 were as follows:

(1) *"February has a much larger amount of jackets sold than the other months shown. From december to january, there was a slight drop in the amount of jackets sold and then a large spike from january to february."*
(2) *"The values in November and May are pretty close, with both being around 37 or 38 jackets. At its peak (February), around 47 jackets were sold."*

This collection was analyzed by a human annotator who manually coded the propositions appearing in each individual summary. This annotation was performed in order to determine, for each graphic, which propositions were used most often.

We developed a set of rules to identify the most important propositions to include in the initial summary of a line graph. Again, due to the continuous nature of time-series data presented in a line graph as opposed to the discrete individual bars in a bar chart, a slightly different approach was necessary. Our content identification rules for line graphs take into account the *magnitude* of a given visual feature (e.g., the *degree* of steepness or volatility) in assigning a weight to the corresponding proposition. In other words, the inclusion of a proposition in the summary of a line graph is not a *yes or no* scenario based on whether or not the measurement of a visual feature exceeds a predefined threshold, but rather, the *strength* of the visual feature affects how much *weight* is assigned to the proposition, and then the highest-ranking propositions are added to the summary.

The weights that a rule assigns to a proposition are based on the relative frequencies of the propositions in the summaries we collected from human subjects for graphs reflecting similar situations. Similar to bar charts, we constructed Message-Category-only rules and Message-Category-Visual-Feature rules.

- Message-Category-only:
  IF $Message - category = m$
  THEN select P with weight w_1

- Message-Category-Visual-Feature:
  IF $Message - category = m$ and $Visual - feature = v$
  THEN select P with weight w_2

However, some propositions (such as conveying the volatility of a line graph) depend only on the visual feature; thus we included a third type of rule, Visual-Feature-only:

- Visual Feature-only:
  IF $Visual - feature = v$
  THEN select P with weight w_2

Message-Category-only rules were constructed when a plurality of human-written summaries in our corpus for all line graphs belonging to a given message category contained the corresponding proposition. Such rules assign a weight according to the frequency with which the proposition was used. As shown in Equation 4, this weight is based on the proportion of summaries for each line graph in the corpus having intended message m and containing proposition P.

$$w_1 = \prod_{i=1}^{n} \frac{P_i}{S_i} \tag{4}$$

In the above equation, n is the number of line graphs in this message category, S_i is the total number of summaries for a particular line graph, and P_i is the number of summaries containing the proposition.

Weights for Message-Category-Visual-Feature and Visual-Feature-only rules are slightly more complicated. The exact definition of this measure depends on the nature of the corresponding visual feature (e.g., steepness of a trend line measured in degrees of angle, volatility of a line), but we normalize all such measures so that they fall in the range from 0 to 1. Since the impact of a visual feature is a matter of degree, the weight assigned by the rule cannot simply rely on the proportion of summaries containing the proposition as in Message-Category-only rules. Instead, we use the covariance between the magnitude of the visual feature ($|v|$) and the frequency of the corresponding proposition ($\frac{P}{S}$) appearing in the corpus summaries for the n graphs displaying this visual feature, as shown in Equation 5:

$$Cov(|v|, \frac{P}{S}) = \left[\left(\frac{\sum_{i=1}^{n} |v_i|}{n} \frac{\sum_{i=1}^{n} \frac{P_i}{S_i}}{n} \right) - \frac{\sum_{i=1}^{n} |v_i| \frac{P_i}{S_i}}{n} \right] \tag{5}$$

Then, for an individual line graph with a magnitude of $|\overline{v}|$ for this visual feature, we compute the weight w_2 for the proposition P as shown in Equation 6:

$$w_2 = |\overline{v}| * Cov(|v|, \frac{P}{S}) \tag{6}$$

Thus, the stronger a certain visual feature is in a given line graph, the higher the weight assigned by the rule for the associated proposition. For example, volatility is one such visual feature; a line graph may represent a trend or changing trend but be very ragged with sharp changes as in Figure 2. To measure the magnitude of this visual feature, our metric takes into account both the frequency of change in a trend and the amplitude of those changes by computing

$$\text{volatility=normalized}(w_1 * w_2)$$

where w_1 is the ratio of the number of short jagged lines comprising the overall segment whose volatility is being measured to the length of the overall segment on the x-axis, and w_2 is the sum of the ratios of the y-axis heights of the short jagged lines comprising the segment to the overall length of the y-axis. The product of w_1 and w_2 is divided by the maximum $w_1 * w_2$ observed for a segment in our corpus, thereby producing a normalized value between 0 and 1. Message-Category-Visual-Feature rules differ from Visual-Feature-only rules in that they are restricted to a specific message category, rather than any line graph exhibiting the particular visual feature.

The line graphs shown in Figures 2 and 10 belong to the same intended message category (Change-trend), and thus would receive the same weights for propositions covered by type 1 rules. However, since Figure 2 is far more volatile than Figure 10, the type 2 rule for volatility would assign a high weight to this proposition for Figure 2, and it is very likely to be included in the initial summary of this line graph (e.g., "*The values in this line graph vary a lot...*"). This proposition might even displace a type 1 proposition that still appears in the summary for Figure 10.

8. ORGANIZING SYSTEM RESPONSES

Once the content of a summary has been determined (i.e., the intended message and important propositions selected), this information must be organized to form a coherent text. This entails deciding how to order propositions with respect to each other and also aggregating individual propositions so that they can be realized as more complex sentences.

We developed a bottom-up generation approach [Demir et al. 2008] for this purpose. Our approach takes advantage of the limited nature of the kinds of relations that can exist between propositions in a descriptive domain [O'Donnell et al. 2001], and utilizes the theory of global focus [Grosz and Sidner 1986] to order propositions. Our approach first groups the propositions into three proposition classes (**message_related, specific,** and **computational**) based on the kind of information they convey about the graph [Demir 2010], and then explores syntactically aggregating propositions within each class. The message_related class contains propositions that convey the intended message of the graphic. The specific class contains the propositions that focus on specific pieces of information in the graphic, such as the proposition conveying the bar with the highest or lowest value. On the other hand, propositions in the computational class capture computations or abstractions over the whole graphic, such as the proposition conveying the overall amount of change in the trend. We hypothesize that the propositions within the message_related class should be presented first in order to emphasize the most important piece of information about the graph. We also hypothesize that the entire graph should be brought back into the user's focus of attention [Grosz and Sidner 1986] before closing the system response. Thus, we define a partial ordering of the proposition classes and present first the message_related propositions, then the specific propositions, and finally the computational propositions.

8.1. Representation

To represent the selected propositions as well as to gain the flexibility necessary to structure and realize a system response, we used a set of basic propositions (minimal information units). The set was defined such that both the intended message and all propositions selected for inclusion can be broken down into these minimal units without information loss. For representing the basic propositions, we defined two kinds of knowledge base predicates (*Relative Knowledge_Base* and *Attributive Knowledge_Base*) and a number of graphical elements, some of which are: 1) *descriptor:* "a referring expression that represents what is being measured in the graphic"[20], 2) *bar(x):* "a particular bar in the graphic", 3) *all_bars:* "all bars depicted in the graphic", and 4) *period(x,y):* "a period depicted in the graphic (between bars x and y)". We used Relative Knowledge_Base predicates to represent the basic propositions which introduce graphical elements or express relations between the graphical elements, and used Attributive Knowledge_Base predicates to represent the basic propositions which present an attribute or a characteristic of a graphical element. Each predicate contains at least two arguments with the first being a graphical element. We refer to the first argument as the "main entity" of a predicate and the other arguments as the "secondary entities" which are either a graphical element or a string constant. For example, consider the graphic in Figure 1 whose primary message is to convey the rank of a bar among the other bars listed (Rank_Bar). Sample instantiations of predicates from the message-related class for the graphic include:

— shows(graphic, bar(3), 3, descriptor, bar_category, all_bars):
"The graphic shows that the third bar is the third highest with respect to the descriptor among all bars of type bar category"

— consists(all_bars, "Visa, Mastercard, American Express, Discover, Diner's Club"):
"The depicted bars (i.e., all bars) are Visa, Mastercard, American Express, Discover, and Diner's Club"

— has_attr(bar_category, "type", "credit card companies"):
"The depicted bars (i.e., all bars) represent credit card companies"

The first two predicates (i.e., shows and consists) are Relative Knowledge_Base predicates and the last predicate (i.e., has_attr) is an Attributive Knowledge_Base predicate.

Each basic proposition (formed by instantiating one of the basic predicates) can be realized as a simple sentence with the main entity in subject position. However, a much more natural text can be generated by aggregating these basic propositions into more complex units. Below we describe a method for aggregating propositions that ensures that the more complex unit can still be realized as a single sentence and balances sentence complexity, readability, and number of sentences in forming a text.

8.2. Aggregation

One strategy for presenting the selected content is to preserve the partial ordering of the proposition classes and to convey each basic proposition via a simple sentence in a predefined order. However, such a straightforward approach would likely yield a text which appeared choppy and disfluent because the focus would be unlikely to flow naturally from one sentence to the next. Thus, we decided to study how basic propositions can be related to each other by aggregating their predicate representations into complex structures. Our aggregation approach treats each basic proposition as a single node tree, and explores different possibilities of obtaining more complex trees (containing multiple propositions) by combining individual trees via four kinds of operators. Our aggregation operators are similar to the clause-combining operations used by the SPoT sentence planner [Walker et al. 2002] in AMELIA. The operators contain applicability constraints which ensure that the complex trees that are formed can be realized as a single sentence (though that sentence may be syntactically quite complex). For a given set of propositions, then, choices exist with respect to how much

[20] Generation of that referring expression using the text associated with the graphic is described in detail in Section 5.

aggregation to perform in order to balance the number of sentences with the syntactic complexity of the individual sentences.

Initially the basic propositions each start as single-node trees. The operators combine trees to form more complex trees. The first three operators (And_Operator, Which_Operator, and Same_Operator) take advantage of opportunities for combining propositions via different types of sentence conjunction (conjoining two sentences with the same subject, forming a relative clause, and conjoining two sentences with the same verb). In addition to combining individual trees, these operators introduce new nodes into the tree structures when they are applied. These new nodes with a single entity correspond to operational predicates And, Which, and Same. The And_Operator and Which_Operator work on trees rooted by a proposition with a Relative Knowledge_Base or an And predicate whereas the Same_Operator works on trees rooted by a proposition with a Relative Knowledge_Base predicate. The last operator, Attribute_Operator, works on propositions with an Attributive Knowledge_Base predicate and essentially recognizes the possibility for a proposition to be realized as an adjective in a noun phrase. Our aggregation operators are defined as follows:

- And_Operator: This operator combines two trees if the propositions at their root share the same main entity. A proposition containing an And predicate with the same main entity forms the root of the new tree and the trees that are combined form the immediate descendents of this root.
- Which_Operator: This operator attaches a tree (Tree T) as a descendent of a node M in another tree (Tree U) via a Which predicate, if the main entity of the proposition at the root of Tree T is a secondary entity for the proposition at node M of the other tree (Tree U). That particular entity forms the main entity of the Which predicate. Thus, Tree T will be an immediate child of the node with the Which predicate and the node with the Which predicate will be an immediate child of node M in Tree U.
- Same_Operator: This operator combines two trees if the propositions at their root contain the same predicate but the main entities of these predicates are different. A proposition with a Same predicate forms the root of the new tree, and the trees that are combined form the descendents of this root. Since the descendents of the new tree have different main entities, the main entity of the Same predicate is some unique element not occurring elsewhere in the tree. For instance, in our implementation this element is obtained by appending a unique number, which isn't used in another Same predicate, to the term "random" (such as "random_0").
- Attribute_Operator: This operator attaches a single node tree that consists solely of a proposition with an Attributive Knowledge_Base predicate, as a direct subchild of a node M with a Relative Knowledge_Base predicate in another tree, if the main entity of the Attributive Knowledge_Base predicate is an entity (main or secondary) for the proposition at node M.

For example, the complex tree structure shown in Figure 22 is produced by applying these operators to the propositions shown in Section 8.1. In this example, the node for the Attributive Knowledge_Base predicate (*) is attached to its parent by the Attribute_Operator which recognized that the main entity (*bar_category*) of the Attributive Knowledge_Base predicate (*has_attr*) is a secondary entity for the *shows* proposition. The node containing the Which predicate (**) is produced by the Which_Operator. This was possible because the main entity (*all_bars*) of the *consists* predicate is a secondary entity of the *shows* proposition. Thus the Which predicate node was introduced to join the original propositions.

We wish to select the best combination of operators to realize each set of basic propositions. Since all aggregation operators cannot be applied to all kinds of predicates and the number of possible sets of basic propositions that should be organized is arguably

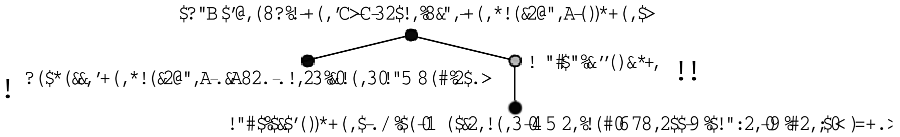


Fig. 22: A complex tree structure.

low, listing all possible combinations of propositions for a proposition class is practical in our work. Thus for each proposition class, we decided to exhaustively generate all possible organizations (i.e., combinations of operators) and determine the best text structure via an evaluation metric (described in Section 8.3). For this purpose, we treat each proposition class (message_related, specific, and computational) as a forest that consists of all single node trees in that class (which we refer to as the "initial candidate forest"). When two trees are combined by an operator, a new candidate forest is produced which is later added to the set of existing candidate forests for that class. Within each proposition class, our approach first applies the And_Operator to all possible pairs of trees in the initial candidate forest, which produces new candidate forests. This is followed by the application of the Same_Operator to all possible pairs of trees in each candidate forest. Similarly, the Which_Operator and finally the Attribute_Operator are applied to trees in the candidate forests produced earlier.

8.3. Evaluation Metric

For each proposition class, our text structuring approach produces a number of candidate forests all of which contain one or more trees with different aggregation (i.e., different realizations). We developed an evaluation metric to determine the best forest which, once realized, would ideally stand at a midpoint between two extremes: a text where each proposition is realized as a single sentence and a text where groups of propositions are realized with sentences that are too complex. Our evaluation metric takes three criteria into account in order to balance these extremes:

$$\text{score}(\mathbf{A}) = \text{nm}_1(\text{sentence}(\mathbf{A})) + \text{nm}_2(\text{complexity}(\mathbf{A})) + \text{nm}_3(\text{clause}(\mathbf{A}))$$

where,[21]

sentence(A): stands for the number of sentences that will be used to realize forest A and equals the number of trees in that forest.
complexity(A): stands for the overall syntactic complexity of forest A and equals the sum of the complexities of sentences that will be used to realize that forest. We used the revised D-level sentence complexity scale [Covington et al. 2006] as the basis of our syntactic complexity measure, and introduced "complexity estimators" to rate individual trees. These are described in [Demir et al. 2008].
clause(A): stands for the overall comprehension complexity of all relative clauses in sentences used for realizing forest A and equals the sum of the comprehension complexities of all clauses. The comprehension complexity of a relative clause equals the product of its syntactic complexity and its position in the sentence, which is equal to 2 if it is a center-embedded clause and is equal to 1 if it is a right-branching clause.

The overall score of a candidate forest is calculated by summing the normalized scores that it receives for each criteria. The scores are normalized with respect to the maxi-

[21] The terms $\mathbf{nm}_1$**,** $\mathbf{nm}_2$**,** and $\mathbf{nm}_3$ stand for the normalized score of a given criteria.

Summary Type	**Best**	**Second-best**	**Third-best**	**Last**
S_O+A+E+	65.6%	26.6%	6.7%	1.1%
S_O+A+E-	16.7%	32.2%	33.3%	17.8%
S_O-A+E+	16.7%	30.0%	40.0%	13.3%
S_O-A-E-	1.0%	11.2%	20.0%	67.8%

Table III: Ranking of Text Structuring and Aggregation

mum score (i.e., 0 <= score <= 1), thus all criteria have an equal impact on the overall score of a forest. Once the scores of all candidate forests are computed, the forest which receives the lowest evaluation score is selected by our approach as the best forest that can be obtained from a set of input propositions. The text structure of a system response consists of the best forests identified for the message_related, specific, and computational classes. Finally, the best text structure is realized via FUF/SURGE [Elhadad and Robin 1999], a publicly available surface realizer.

8.4. Evaluation of the Text Structuring and Aggregation Approach

We conducted a user study to evaluate whether or not our decisions with respect to different components of our approach contribute to the quality of the generated text. The components we evaluated were: i) the organization and ordering **(O)** of the content (partial ordering of the proposition classes and classification of the propositions), ii) the aggregation **(A)** of the selected content into candidate forests, and iii) the metric **(E)** used to evaluate candidate forests.

Fifteen university undergraduate or graduate students participated in the study. Each evaluator was presented with four different summaries of each of twelve graphics from our corpus. The participants did not participate in any of our earlier user studies nor were they involved in this work. Although all summaries that we presented to the participants were automatically produced by our generation approach, the participants were unaware of this fact (i.e., whether summaries were human-generated or computer-generated). In the study, we only used graphics that conveyed an increasing or a decreasing trend, since these message categories exhibited the greatest variety of possible summaries. For each graphic, the participants were asked to rank the given set of summaries based on their quality in conveying the content. The summaries varied according to the test parameters as follows:

— **S_O+A+E+:** A summary that uses the ordering rules, the aggregation rules, and receives the lowest (best) overall score by the evaluation metric. This is the summary selected as best by our approach.
— **S_O+A+E-:** A summary that uses the ordering and aggregation rules, but does not receive the lowest overall score by the evaluation metric. This is the summary that received the second lowest score.
— **S_O-A+E+:** A summary where the propositions are randomly ordered (as opposed to being grouped into message_related, specific, and computational classes), but aggregation takes place, and it receives the lowest overall score by the evaluation metric.
— **S_O-A-E-:** A summary consisting of single sentences that are randomly ordered.

Table III displays the results of the evaluation. The results of the experiment revealed that the summary selected as the best by our approach was most often (**65.6%** of the time) rated as the best summary by the participants. Our approach was also shown to select the summary which was rated as one of the top two summaries **92.2%** of the time. Moreover, the study showed that the summaries where the evaluation metric (S_O+A+E-) or the ordering of propositions (S_O-A+E+) was omitted were substantially less preferred by the participants. Overall, this user study validated our text structuring and aggregation approach.

8.5. Evaluating Consistency of Bar Chart Initial Summary Content

To determine whether the knowledge that users were gleaning from our summaries was consistent with the content of the graphics, we performed an experiment in which 18 subjects were presented with our system's initial summary for each of four bar charts. The selected bar charts conveyed an increasing trend, a decreasing trend, the entity with the greatest value, and the rank of a particular entity among all the entities depicted by bars in the bar chart. The subjects were then asked to draw the bar chart described by the summary.

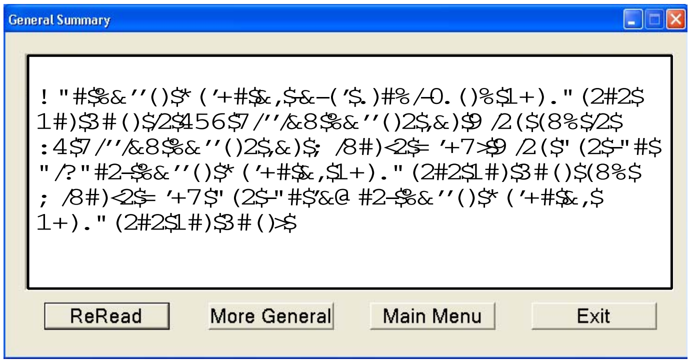
Three evaluators rated 68 drawings[22] on a scale of 1 to 5, with 5 being the best rating and indicating that the drawing captured the most important information that should be conveyed about the bar chart while a rating of 1 indicated that the drawing failed to reflect the original graphic. The Fleiss kappa for inter-rater agreement is close to 0. However, we believe that the Kappa statistic does not appropriately measure agreement on rating tasks, since scores of 5 and 4 for a graph are treated the same as scores of 5 and 1, yet the former reflects closer agreement about the quality of the graph than does the latter. Thus we used Kendall's W (also known as Kendall's coefficient of concordance) [Kendall and Babington-Smith 1939; Daniel 1989]. The computed W value is .507. For large samples (more than 7 drawings), one computes a value (based on W) to be used in a χ^2 table which for our experiment is 101.86 with 67 degrees of freedom. Thus at the .01 significance level, we can reject the null hypothesis and conclude that there is agreement among the evaluators.

The graphs conveying an increasing trend and a decreasing trend received average scores of 4.22 and 4.63 respectively with standard deviations of .94 and .55. The graphs conveying the entity with the maximum value and the rank of a particular entity received average scores of 3.53 and 4.07 respectively with standard deviations of .98 and .84. Since our initial summaries do not contain all features of a bar chart (such as the value for every bar), it is not surprising that the overall average score is only 4.11. We collected feedback from the evaluators regarding the reasons for their ratings. In the case of graphs presenting an increasing or decreasing trend, missing values on the dependent axis and missing measurement axis descriptors (although they were given in the summary) were the main reasons for lower ratings. We argue that accurate y-axis labels are more important for scientific graphs than for graphs from popular media where the key to understanding the article is assimilation of the high-level content of the graphic. In the case of graphs conveying the entity with the maximum value or the rank of a particular entity, the main reason for lower ratings was incorrect ranking of some bars in the drawings. Our summaries do not convey the ranks and values of all bars for these two kinds of graphs, and we argue that the rank of every bar in the graph is not essential to understanding the high-level content of the graph and thus should not be presented in the initial summary.

9. GENERATING FOLLOW-UP RESPONSES

An information graphic depicts a large amount of information. Throughout the development of the system, we observed that users have different preferences for the amount and kind of information that they want to receive about particular graphics from popular media. Often users were only interested in the high-level content of a graphic, whereas at other times users wanted to receive more detailed information

[22] Four drawings were discarded because the subject drew a pie chart instead of a bar chart



```
! "#$%&''()$* ('+#$&,$&–('$.)#%/–0.()%$1+)." (2#2$
1#)$3#()$/2$456$7/''/&8$%&''()2$,&)$9 /2($(8%$/2$
:4$7/''/&8$%&''()2$,&)$; /8#)<2$= '+7>$9 /2($' (2$–"#$
"/?"#2–$%&''()$* ('+#$&,$1+)." (2#2$1#)$3#()$(8%$
; /8#)<2$= '+7$' (2$–"#$'&@ #2–$%&''()$* ('+#$&,$
1+)." (2#2$1#)$3#()>$
```

Fig. 23: General follow-up response after the initial summary.

about a graph (such as individual bar values) perhaps because they were more interested in the subject the graphic was about. Presenting everything about a graphic at once is not practical since this would overwhelm those users who might not be interested in many details about the graph. On the other hand, presenting only the high-level content would not satisfy users who have specific points of interest about the graphic. Our system presents only a subset of the identified propositions in the initial summary. Thus, there remains many unconveyed propositions which can be made available to the users in order to allow them to receive as much information as they desire about the graphic. Unfortunately, human subject experiments [Schwartz 2007] revealed a major problem for congenitally blind participants in that they could not identify what other kinds of information they might request after receiving the initial summary of a graphic. We therefore provided a follow-up facility via a keystroke-driven menu-based interface in order to enable users to glean more in-depth information about a graph after receiving its initial summary. This follow-up mechanism also eliminates possible speech recognition problems and issues in interpreting free-form questions. Our follow-up facility is fully implemented for bar charts but extending the overall methodology to other kinds of information graphics such as line graphs is in our future research agenda. In the system, we offer three different kinds of follow-up responses and present them as options in a dialog window (as shown in Figure 4):

- **General Follow-up Response:** This option is for users who want to receive additional information about the graphic without specifying any preferences. If this follow-up response is requested by a user, our system first ranks and then selects the most highly-rated propositions from among the propositions that could be conveyed about the graphic. For example, for the graphic in Figure 1, if the user indicates his desire for a general follow-up response by clicking the "General Information" button or pressing "SHIFT+G" in the dialog window shown in Figure 4, our system generates the follow-up response shown in Figure 23.
- **Focused Follow-up Response:** This option is for users who desire a particular kind of further information about the graphic (such as more information about the change observed in a trend). For each intended message category, we identified the relevant propositions and classified these propositions into a number of different categories (up to five categories). If this follow-up response is requested by a user (by clicking the "Focused Information" button in the main follow-up window shown in Figure 4 or by pressing "SHIFT+F"), the system first prompts the user to select the follow-up category that he is interested in. For example, for graphics that convey the rank of a bar, our system offers the categories shown in Figure 24: i) a comparison

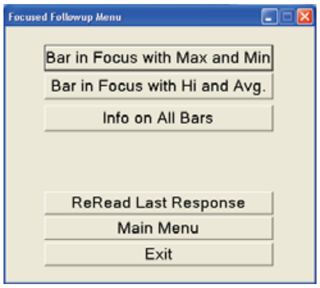
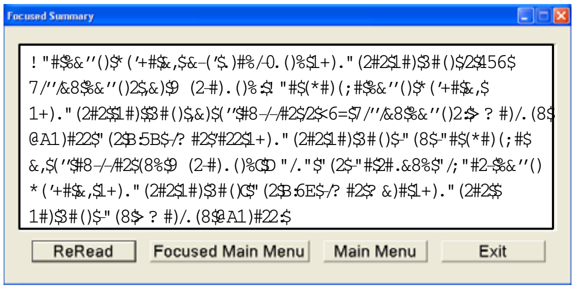


Fig. 24: Main focused follow-up menu.



```
!"#$%&''()$*('+#$&,$&-('$.)#%/-0.()%$1+)."(2#2$1#)$3#()$/2$456$
7/''/&8$%&''()2$,&)$9 (2-#).()%:$! "#$(*#)(;#$%&''()$*('+#$&,$
1+)."(2#2$$1#)$$3#()$,&)$(''$#8-/-/#2$/2$<6=$7/''/&8$%&''()2:$>?#)/.(8$
@A1)#22$"(2$B:5B$-/? #2$'#22$1+)."(2#2$1#)$3#()$-"(8$-"#$(*#)(;#$
&,$(''$#8-/-/#2$(8%$9 (2-#).()%C$D"/."$"(2$-"#$2#.&8%$"/;"#2-$%&''()
*('+#$&,$1+)."(2#2$1#)$3#()C$"(2$B:6E$-/? #2$? &)#$1+)."(2#2$
1#)$3#()$-"(8$>?#)/.(8$@A1)#22:$
```

Fig. 25: Focused follow-up response after the summary.

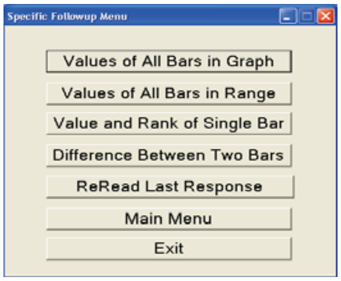
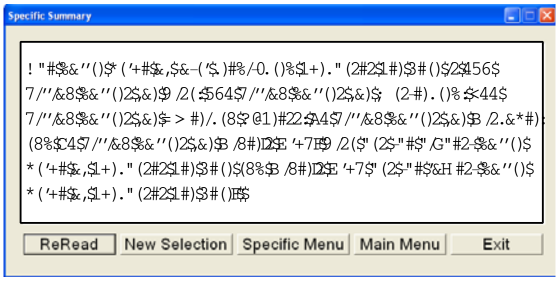

Specific Followup Menu
Values of All Bars in Graph
Values of All Bars in Range
Value and Rank of Single Bar
Difference Between Two Bars
ReRead Last Response
Main Menu
Exit

Fig. 26: Main specific follow-up menu.



```
!"#$%&''()$*('+#$&,$&-('$.)#%/-0.()%$1+)."(2#2$1#)$3#()$/2$456$
7/''/&8$%&''()2$,&)$9 /2(:$564$7/''/&8$%&''()2$,&)$; (2-#).()%:$<44$
7/''/&8$%&''()2$,&)$=>#)/.(8$?@1)#22:$A4$7/''/&8$%&''()2$,&)$B /2.&*#):
(8%$C4$7/''/&8$%&''()2$,&)$B /8#)D2$E'+7E$9 /2($"(2$-"#$"/G"#2-$%&''()$
*('+#$&,$1+)."(2#2$1#)$3#()$(8%$B /8#)D2$E'+7$"(2$-"#$'&H#2-$%&''()$
*('+#$&,$1+)."(2#2$1#)$3#()E$
```

Fig. 27: Specific follow-up response after the summary.

of the bar in focus with the bars that have the highest and the lowest values, ii) a comparison of the bar in focus with the bar that has the closest higher value, and a comparison of the value of the bar in focus with the average value for all bars, and iii) more information about all bars listed in the graphic. Our system then ranks the propositions and presents the most highly-rated propositions classified in the selected category. For example, suppose that after receiving the initial summary for the graphic in Figure 1, the user selects “Focused Information” from the main follow-up menu in Figure 4 and then selects the second focused follow-up category (*Bar in Focus with Hi and Avg*) from the focused follow-up menu in Figure 24; our system then produces the follow-up response shown in Figure 25.

- **Specific Follow-up Response:** This option is for users who request specific information about the graphic (i.e., those aspects of the graph that he finds interesting) such as the rank of a specific bar in a bar chart. The information of interest to the user might not be directly related to the graphic’s intended message or might not be conveyed in the other kinds of follow-up responses or the initial summary. If this follow-up response is requested by the user (by clicking the “Specific Information” button or pressing “SHIFT+S” in the main follow-up window), specific follow-up categories that are common to all intended message types are presented to the user (as shown in Figure 26). For example, for the graphic in Figure 1, if the user selects the first option from the main specific follow-up menu after receiving its initial summary, the follow-up response shown in Figure 27 is presented by the system. No ranking takes place in this kind of follow-up response, but the response might contain additional information related to what is requested by the user.

9.1. Content Selection for Follow-up Responses

To determine the content of general and focused follow-up responses, we developed a sophisticated graph-based ranking approach [Demir et al. 2010]. Our approach ranks a set of propositions with respect to their importance at this point in the interaction with the user so that the most highly-rated of these propositions can be conveyed in a follow-up response. We leveraged three main considerations in this approach in order to produce coherent responses: i) unrelated propositions should not be conveyed in the same response, ii) propositions that are normally communicated together should appear in the same response, and iii) redundant information should be avoided in a response. Our ranking approach utilizes an undirected weighted graph (which we refer to as the "relation graph") where all identified propositions are represented as vertices and the relations between propositions are represented as edges. For each intended message category, a different relation graph is constructed since the set of applicable propositions varies between these categories. For example, one message category for bar charts is *Maximum* (conveying the entity whose value is greatest). The proposition capturing the value of the second largest entity is applicable for graphs whose intended message falls into this message category but is not applicable if the message category is *Falling-trend*. Four relation classes are defined to classify relations between propositions where each is assigned a different numeric score indicating how important it is to convey the propositions sharing that relation in the same follow-up response. These scores are, in turn, used to specify weights on the edges of the graph. In order to asses the a priori importance of the propositions, a stereotypical user model is incorporated in this graph-based setting.

In this representation, our task is to determine the importance of each vertex and rank these vertices accordingly. Our ranking approach iteratively runs the weighted PageRank[23] [Brin and Page 1998; Sinha and Mihalcea 2007] and makes proper adjustments on the edges of the relation graph after each iteration. For instance, the PageRank algorithm is run and the proposition that is deemed most important is selected for inclusion in the follow-up response. Then the weights on the edges connected to this proposition are updated so that propositions that should be communicated with it are given higher weight, and propositions that are redundant with it are given a lower rate. Then the algorithm is run again to select the next proposition and the cycle is repeated. These adjustments are made to achieve the design considerations behind our approach, namely to favor the selection of related propositions in the same response and to discourage the selection of propositions that are unrelated or that convey redundant information.

Our content determination approach has two other notable features which are necessary for modeling natural human interactions: i) generating successive history-aware responses and ii) being flexible to generate different responses with different parameter settings (e.g., generating user-tailored responses given a user model). For example, our approach keeps a dialogue history in order to relate the content of the current response to previously communicated propositions (i.e., in the initial summary and previous follow-up responses of any kind). This enables the system to omit information that has been recently conveyed, to determine when repetition of previously communicated information is appropriate, and to use discourse markers to signal these repetitions.

Our content selection methodology was evaluated via two user studies where our focus was on the content of general follow-up responses.[24] In both studies, bar charts

[23] This metric calculates the importance of a vertex by taking into account the importance of all other vertices and the relation of vertices to one another.

[24] Focused and specific followup responses are implemented in the SIGHT system but have not yet been evaluated.

Graph	Number of raters	Average rating	Standard deviation
1	9	3.44	.96
2	3	3.67	.94
3	2	4.00	0.00
4	3	4.67	.47
5	4	4.25	.83
6	2	5.00	0.00
7	4	2.75	1.48
8	5	3.2	.75
9	7	3.71	1.03
10	10	3.6	1.28
11	3	4.17	.24
12	5	3.6	.80

Table IV: Statistics for each graph that was rated

with an increasing or a decreasing trend were used. Before the studies, each participant was informed that they would be first given graphics along with their initial summaries which should contain the most important information about the graph. They were also told that the remaining pieces of information would be conveyed via general follow-up responses where the information given in the first follow-up response should be more important than the information in the second follow-up response. In the first study, nineteen graduate students were presented with three graphics randomly selected from twelve test graphics along with the initial summary.[25] Then, they were presented with two consecutive follow-up responses. Each participant was asked to evaluate the system responses on the basis of their satisfaction (on a scale of 1 to 5 with 5 being the best) and to specify whether or not any part of a response should be communicated earlier or omitted. Table IV displays the number of subjects who rated each graph, the average rating, and the standard deviation. Overall, the participants rated the system favorably with an average of **3.69**, a standard deviation of **1.08**, and a median of **4**. In addition, no consensus was observed among the participants in regard to what content should be added or eliminated in the system responses.

In the second study, twenty one university students who did not participate in the first study were presented with the same four graphics (selected from the twelve test graphics described above). For each graphic, the participants were first presented with its initial summary and a set of 18 propositions that were used to construct the relation graph for graphics with an increasing or a decreasing trend. Each participant was then asked to select four propositions that they deemed most important to be conveyed in the first general follow-up response.

Surprisingly, the number of times that a proposition was selected by a participant ranged from 0 to 13, with all but one or two propositions in every graph selected by at least one participant. For the four graphs, the most popular proposition was selected by 13, 11, 13, and 12 of the participants respectively and the second most popular proposition by 12, 11, 13 and 9 of the participants respectively. If every proposition were selected the same number of times, each proposition would be selected 4.67 times. The number of propositions whose rate of selection exceeded this average was 8, 8, 7, and 6 for the four graphs. These statistics show that while there was some consensus about

[25] Although Fleiss's Kappa does not require that each graphic be rated by the same subjects, it does require that each graphic receive the same number of ratings. Since each subject was given a random selection of three graphics, the number of ratings per graphic varies. Thus a Kappa statistic of inter-rater agreement cannot be computed.

the best proposition to include and some agreement about the second best proposition, the selection of subsequent propositions was subjective and varied greatly.

For three of the four graphics, our system included the most popular proposition in its followup response and for two of the graphs, our system included the second most popular proposition. For all of the graphs, the system included three propositions from among the six most popular propositions for that graphic, and in two cases included four propositions from among the six most popular ones. Furthermore, the average number of times that the participants selected the four most popular propositions was 9.9 whereas the average number of times that the participants selected the same propositions selected by the system was 7.8. This study showed that the system was 1) successful at identifying a very important proposition to include in the followup response, 2) with the exception of one graphic, all of the propositions selected by the system were viewed as important (i.e., they ranked among the top six in terms of number of participants who selected them), and 3) the average number of times that the system's propositions were also selected by the participants was well above the average of 4.67. Thus from this study and the previous experiment, we conclude that the system produces high-quality followup responses. However, future work will examine how this content selection might be improved.

10. PRELIMINARY EVALUATION OF THE SIGHT SYSTEM

The goal of the SIGHT system is to provide blind individuals (or those with seriously impaired eyesight) with access to information graphics in popular media. To evaluate the effectiveness of our system, we conducted human subjects experiments in which the subjects used SIGHT to answer important questions whose answers could be gleaned only from a set of information graphics. To identify the important questions, referred to as *key* questions, we first performed a two-stage experiment with sighted subjects. In the first stage, 17 subjects were given a set of twelve bar charts and asked to list the three most important questions whose answers they believed one should glean from the graphic. The subjects had no knowledge of the SIGHT system or the kinds of information it could provide. To ensure that the questions could be answered from the graphic alone, we conducted a second stage of the experiment in which two graduate students in computational linguistics analyzed the questions and removed any that did not meet this restriction. We then ranked the questions by the number of times a question was identified by a subject in the first stage of the experiment, and labeled the questions that appeared most often for graphics in a particular message category as *key* questions — that is, questions that our SIGHT system should certainly facilitate answering. For example, for graphics whose intended message falls into the message category *Rank* (conveying the rank of an entity with respect to other entities), the following are three representative key questions:

— What is the value of the bar whose rank is in focus in the graphic?
— Which bar has the highest value?
— What is the value of the lowest bar?

We then recruited seven people with significant visual impairments, and who were familiar with using a computer to read text either with a screenreader or a screen magnifier, for a quantitative and qualitative evaluation of the SIGHT system. Some were totally blind and the others had impaired eyesight to the extent that they used a screen magnifier and thus could only view small pieces of a graphic at once. Each of these subjects was shown how to use the SIGHT system and allowed to practice with it. They were then given three graphics whose intended messages fell into different message categories, and presented with the task of using SIGHT to obtain the answers to three randomly selected key questions for that graphic. After the subjects completed the tasks, an interview was conducted regarding the usability and effectiveness of the SIGHT system.

The subjects successfully answered all of the key questions. Occasionally, the subjects used only the initial summary of the graphic but in most cases they requested followup information. The kinds and order of followup requests varied among the subjects, but all three kinds of followup (general, focused, and specific) were utilized by the group. These results led us to conclude that 1) SIGHT is effective at providing access to the kind of information that sighted subjects identify as important to glean from an information graphic in popular media, and 2) the three different kinds of followup information serve a role in providing appropriate access to information graphics.

During the interview following the experiment, the subjects were asked to rate the system on a scale from 1 to 10 with respect to usefulness and ease of use, where high numbers equated with being extremely useful or very easy to use. The average rating was 9.43 (standard deviation .73) for usefulness and 8.71 (standard deviation .88) for ease of use. The subjects were overwhelmingly positive about the system and the flexibility provided by the different kinds of followup responses. All of the subjects stated that they would use SIGHT if it were available for everyday web usage. As stated by one subject,

> *"I think that having a system that can describe bar charts to blind and visually impaired users is an extremely valuable resource. If this program had been available to me, I would have had the ability to function as everyone else would."*

From these studies, we conclude that SIGHT's methodology is effective at providing access to the knowledge that one should glean from an information graphic. However our visually impaired subjects mentioned three limitations of the system that must be addressed in future work. Although various accessibility recommendations were taken into account while designing the user interface, individuals who use screen magnification software found that some layout elements (e.g., radio buttons in some specific follow-up dialog windows) were difficult for individuals to find and click on. The layout should be improved to make these elements stand out more from the background. Second, pressing keyboard accelerators or clicking navigational buttons to interact with the system costs more time and effort for visually impaired users who also have moderate to severe motor impairments. Thus, integrating speech recognition into the interface would engage those users more easily and comfortably in the system. Finally, the time that it takes to generate a system response in some message categories (i.e., those that require comparing a bar with others such as the Maximum Bar message category) is significantly longer than what is expected by the users. Therefore, techniques for parallelizing system execution should be investigated to decrease the response time.

In the future, we will conduct more extensive experiments in which we compare knowledge acquisition from multimodal documents by blind subjects with and without the access to information graphics provided by SIGHT. In addition, we will perform experiments in which blind subjects comment on what and why they are choosing different kinds of followup in order to find desired information. Other colleagues are developing a system that generates tactile representations of graphics from electronic documents. We anticipate further experiments in which the two approaches are compared, the advantages and disadvantages of each are identified, and a system that offers both in combination will be investigated.

11. FUTURE WORK

Our overall system evaluation has been concerned with whether SIGHT provides users with easy access to the important knowledge conveyed by information graphics. Future evaluation studies will examine the effectiveness of SIGHT in providing appropriate access to information graphics in the context of reading a multimodal document. We will invite individuals who are blind to use our system extensively while perusing articles in popular media, and we will solicit in-depth feedback about their experiences in order to identify modifications to our interface or methodology that will improve the user experience.

In addition to extending SIGHT to handle pie charts, grouped bar charts [Burns et al. 2010; Burns et al. 2012], and composite graphics, we plan to explore how we can provide a better overall experience for the user through deeper integration of the text and graphical content. Currently, the information graphic is analyzed and summarized *in isolation*, and its description is simply inserted into the most relevant paragraph without any connection with or transition to/from the textual content. We want to improve this situation in three ways: taking the article text into account when building a summary of the graphic, presenting the graphical summary in a more cohesive way with respect to the article text, and ultimately generating an abstractive summary of the entire multimodal document covering both the text and graphical content.

Employing a novel architecture for abstractive summarization [Greenbacker et al. 2011], we will represent the entire multimodal document in a single, unified semantic model. The intended message and other salient propositions extracted from the information graphics by SIGHT will be decomposed and added to a model of the text built with a semantic parser using a domain-specific grammar. Working from this unified semantic model will enable our system to take both the text and graphical content into account when producing a summary of the graphic. Rather than selecting propositions from the information graphic based solely on its intended message and visual features, the focus of the text can influence the choice of propositions. For example, if a line graph is only moderately volatile, the corresponding proposition may not be weighted highly enough to warrant inclusion in the initial summary. However, the article text might include an emphasis on the volatility of the data presented in the graphic (e.g., by describing the "ups & downs" of a company's stock price). The semantic model will allow us to recognize this special focus and use it to boost the weight of the proposition for volatility so that it appears in the summary of the line graph. Additionally, by understanding the structure and content of the article text, we will be able to determine a more specific location within a paragraph at which to include the summary of the graphic, and adjust the exact wording of the summary to ensure a natural transition and improve cohesion.

Another benefit of working from a semantic model of the text and graphics is that we will be able to generate a unified summary of the entire multimodal document, covering content extracted from both sources. Our goal will be an interactive system where the user can choose to access an entire multimodal document or alternatively can request a summary whose length is a user-specified percent of the document's length. Such output will provide individuals having sight impairments with the option to evaluate a summary when deciding whether or not to commit the time required to read a lengthy article, or to quickly get a sense of the most important concepts without having to read the entire document.

12. CONCLUSION

SIGHT is an intelligent interactive system for providing blind individuals with access to the knowledge conveyed by information graphics in popular media. Rather than attempting to provide access to the graphic's appearance by rendering it in an alternative medium such as sound or touch or by using speech to describe the appearance of the graphic, SIGHT conveys the knowledge that one would glean from viewing the graphic. It uses the JAWS screenreader within a web browser to communicate information to the user via speech, provides the user with the opportunity to obtain a brief summary of a graphic's high-level content at the most relevant point in a multimodal article, and enables the user to obtain more detailed followup information if desired. Evaluation of SIGHT with sight-impaired individuals suggests that SIGHT is effective at providing access to information graphics. Thus SIGHT significantly advances the opportunity for universal access to multimodal documents. Future work will focus on extending SIGHT's capability, particularly with regard to handling more complex information graphics and to enabling the user to obtain multimodal document summaries of user-specified length.

REFERENCES

ABU DOUSH, I., PONTELLI, E., SIMON, D., AND MA, O. 2009. Making Microsoft Excel accessible: Multimodal presentation of charts. In *Proceedings of the Eleventh International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS 2009. ACM, Pittsburgh.

ALM, N., NEWELL, A., AND ARNOTT, J. 1987. A communication aid which models conversational patterns. In *Proceedings of the Tenth Annual Conference on Rehabilitation Technology*, R. Steele and W. Gerrey, Eds. RESNA, Washington, DC, 127–129.

ALTY, J. L. AND RIGAS, D. 2005. Exploring the use of structured musical stimuli to communicate simple diagrams: the role of context. *International Journal of Human-Compututer Studies 62,* 1, 21–40.

ANG, J., DHILLON, R., KRUPSKI, A., SHRIBERG, E., AND STOLCKE, A. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of the 7th International Conference on Spoken Language Processing*. ICSLP-2002. ISCA, Denver, 2037–2040.

BLACK, R., REDDINGTON, J., REITER, E., TINTAREV, N., AND WALLER, A. 2010. Using NLG and sensors to support personal narrative for children with complex communication needs. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*. ACL, Los Angeles, 1–9.

BRADLEY, D. C., STEIL, G. M., AND BERGMAN, R. N. 1995. OOPSEG: a data smoothing program for quantitation and isolation of random measurement error. *Computer Methods and Programs in Biomedicine 46,* 1, 67–77.

BRENNAN, R. AND PREDIGER, D. 1981. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement 41,* 3, 687–699.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems 30,* 1–7, 107–117.

BROWN, L. M. AND BREWSTER, S. A. 2003. Drawing by ear: interpreting sonified line graphs. In *Proceedings of the 9th Meeting of the ICAD: International Conference on Auditory Display*. ICAD03. ICAD, Boston, 152–156.

BURNS, R., CARBERRY, S., AND ELZER, S. 2010. Visual and spatial factors in a Bayesian reasoning framework for the recognition of intended messages in grouped bar charts. In *Proceedings of the AAAI 2010 Workshop on Visual Representations and Reasoning*. AAAI Press, Atlanta, 6–13.

BURNS, R., CARBERRY, S., ELZER, S., AND CHESTER, D. 2012. Automatically recognizing intended messages in grouped bar charts. In *Proceedings of the 7th International Conference on the Theory and Application of Diagrams*. To Appear.

CARBERRY, S. AND ELZER, S. 2007. Exploiting evidence analysis in plan recognition. In *the Proceedings of the International Conference on User Modeling*.

CARBERRY, S., ELZER, S., AND DEMIR, S. 2006. Information graphics: an untapped resource for digital libraries. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '06. ACM, Seattle, 581–588.

CHEN, C. 2004. *Information visualization: Beyond the Horizon*. Springer.

CLARK, H. H. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.

COCKBURN, A., KARLSON, A., AND BEDERSON, B. B. 2008. A review of overview+detail, zooming, and focus+context interfaces. *ACM Computing Surveys 41*, 1–31.

COHEN, J. A. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20,* 1, 37–46.

COHEN, R. F., YU, R., MEACHAM, A., AND SKAFF, J. 2005. Plumb: displaying graphs to the blind using an active auditory interface. In *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS 2005. ACM, Baltimore, 182–183.

COVINGTON, M. A., HE, C., BROWN, C., NAÇI, L., AND BROWN, J. 2006. How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-level scale. CASPR Research Report 2006-01, Artificial Intelligence Center, University of Georgia, Athens, Georgia.

COYNE, P. AND NIELSEN, J. 2001. *Beyond ALT Text: Making the Web Easy to Use for Users with Disabilities*. Nielsen-Norman Group, Fremont, California.

DANIEL, W. 1989. *Applied Nonparametric Statistics: Second Edition*. PWS-Kent Publishing.

DEMIR, S. 2010. SIGHT for visually impaired users: Summarizing information graphics textually. Ph.D. thesis, University of Delaware.

DEMIR, S., CARBERRY, S., AND MCCOY, K. F. 2008. Generating textual summaries of bar charts. In *Proceedings of the 5th International Natural Language Generation Conference*. INLG 2008. ACL, Salt Fork, Ohio, 7–15.

DEMIR, S., CARBERRY, S., AND MCCOY, K. F. 2010. A discourse-aware graph-based content-selection framework. In *Proceedings of the 6th International Natural Language Generation Conference*. INLG 2010. ACL, Trim, Ireland, 17–25.

ELHADAD, M. AND ROBIN, J. 1999. SURGE: a comprehensive plug-in syntactic realization component for text generation. Technical report, Department of Computer Science, Ben-Gurion University of the Negev, Beer Sheva, Israel.

ELZER, S. 2005. A probabilistic framework for the recognition of intention in information graphics. Ph.D. thesis, University of Delaware.

ELZER, S., CARBERRY, S., CHESTER, D., DEMIR, S., GREEN, N., ZUKERMAN, I., AND TRNKA, K. 2005. Exploring and exploiting the limited utility of captions in recognizing intention in information graphics. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. ACL, Ann Arbor, 223–230.

ELZER, S., CARBERRY, S., ZUKERMAN, I., CHESTER, D., GREEN, N., AND DEMIR, S. 2005. A probabilistic framework for recognizing intention in information graphics. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 1042–1047.

ELZER, S., GREEN, N., CARBERRY, S., AND HOFFMAN, J. 2006. A model of perceptual task effort for bar charts and its role in recognizing intention. *User Modeling and User-Adapted Interaction 16*, 1–30. Received James Chen best paper award.

FERRES, L., LINDGAARD, G., AND SUMEGI, L. 2010. Evaluating a tool for improving accessibility to charts and graphs. In *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS '10. ACM, Orlando, 83–90.

FERRES, L., VERKHOGLIAD, P., LINDGAARD, G., BOUCHER, L., CHRETIEN, A., AND LACHANCE, M. 2007. Improving accessibility to statistical graphs: the iGraph-Lite system. In *Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS 2007. ACM, Tempe, Arizona, 67–74.

FLEISS, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin 76,* 5, 378–382.

FLOWERS, J. H. AND HAUER, T. A. 1995. Musical versus visual graphs: Cross-modal equivalence in perception of time series data. *Human Factors: The Journal of the Human Factors and Ergonomics Society 37*, 553–569.

FRITZ, J. P. AND BARNER, K. E. 1999. Design of a haptic data visualization system for people with visual impairments. *IEEE Transactions on Rehabilitation Engineering 7*, 372–384.

GERBER, E. 2002. Surfing by ear: Usability concerns of computer users who are blind or visually impaired. *AccessWorld 3,* 1, 38–43.

GIPS, J. 1998. On building intelligence into EagleEyes. In *Assistive Technology and Artificial Intelligence, Applications in Robotics, User Interfaces and Natural Language Processing*, V. O. Mittal, H. A. Yanco, J. Aronis, and R. Simpson, Eds. Springer-Verlag, London, 50–58.

GONCU, C. AND MARRIOTT, K. 2008. Tactile chart generation tool. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS 2008. ACM, Halifax, Nova Scotia, 255–256.

GONCU, C., MARRIOTT, K., AND HURST, J. 2010. Usability of accessible bar charts. In *Proceedings of the Sixth International Conference on the Theory and Application of Diagrams*. Diagrams 2010. Springer-Verlag, Portland, Oregon, 167–181.

GREEN, N. L., CARENINI, G., KERPEDJIEV, S., MATTIS, J., MOORE, J. D., AND ROTH, S. F. 2004. Autobrief: an experimental system for the automatic generation of briefings in integrated text and information graphics. *International Journal of Human Computer Studies 61,* 1, 32–70.

GREENBACKER, C. F., MCCOY, K. F., CARBERRY, S., AND MCDONALD, D. D. 2011. Semantic modeling of multimodal documents for abstractive summarization. In *Proceedings of the Canadian AI 2011 Workshop on Automatic Text Summarization*. University of Ottawa, St. John's, Newfoundland, 29–40.

GRICE, H. P. 1975. Logic and Conversation. In *Syntax and Semantics III: Speech Acts*, P. Cole and J. L. Morgan, Eds. Academic Press, N.Y., 41–58.

GROSZ, B. AND SIDNER, C. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics 12,* 3, 175–204.

INA, S. 1996. Computer graphics for the blind. *SIGCAPH Computers and the Physically Handicapped 55*, 16–23.

JAYANT, C., RENZELMANN, M., WEN, D., KRISNANDI, S., LADNER, R., AND COMDEN, D. 2007. Automated tactile graphics translation: in the field. In *Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS 2007. ACM, Tempe, Arizona, 75–82.

JOHN, B. E. AND NEWELL, A. 1990. Toward an engineering model of stimulus response compatibility. In *Stimulus-response compatibility: An integrated approach*, R. W. Gilmore and T. G. Reeve, Eds. North-Holland, New York, 107–115.

JOSHI, A., WEBBER, B., AND WEISCHEDEL, R. M. 1986. Living up to expectations: computing expert responses. In *Proceedings of the Strategic Computing - Natural Language Workshop*. HLT '86. ACL, Marina del Rey, California, 179–189.

KENDALL, M. AND BABINGTON-SMITH, B. 1939. The problem of m rankings. *The Annals of Mathematical Statistics 10,* 3, 275–287.

KENNEL, A. R. 1996. Audiograf: a diagram-reader for the blind. In *Proceedings of the Second Annual ACM Conference on Assistive Technologies*. ASSETS 1996. ACM, Vancouver, British Columbia, 51–56.

KRUFKA, S. E. AND BARNER, K. E. 2006. A user study on tactile graphic generation methods. *Behaviour and Information Technology 25,* 4, 297–311.

KURZE, M. 1995. Giving blind people access to graphics (example: Business graphics). In *Proceedings of the Software-Ergonomie '95 Workshop on Nicht-visuelle graphische Benutzungsoberflächen (Non-visual Graphical User Interfaces)*. B.G. Teubner, Darmstadt, Germany.

LARKIN, J. AND SIMON, H. 1987. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science 11*, 65–99.

LEWANDOWSKY, S. AND SPENCE, I. 1989. The perception of statistical graphs. *Sociological Methods and Research 18,* 2 & 3, 200–242.

LIU, X. AND CROFT, W. B. 2002. Passage retrieval based on language models. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*. CIKM '02. ACM, McLean, Virginia, 375–382.

LOHSE, G. L. 1993. A cognitive model for understanding graphical perception. *Human-Computer Interaction 8*, 353–388.

MACAULAY, F., JUDSON, A., ETCHELS, M., ASHRAF, S., RICKETTS, I. W., WALLER, A., BRODIE, J. K., ALM, N., WARDEN, A., SHEARER, A. J., AND GORDON, B. 2002. ICU-Talk, a communication aid for intubated intensive care patients. In *Proceedings of the Fifth International ACM Conference on Assistive Technologies*. ASSETS 2002. ACM, Edinburgh, Scotland, 226–230.

MCCOY, K. F., CARBERRY, M. S., ROPER, T., AND GREEN, N. 2001. Towards generating textual summaries of graphs. In *Proceedings of the 1st International Conference on Universal Access in Human-Computer Interaction*. UAHCI 2001. Lawrence Erlbaum, New Orleans, 695–699.

MCGOOKIN, D. K. AND BREWSTER, S. A. 2006. Soundbar: exploiting multiple views in multimodal graph browsing. In *Proceedings of the 4th Nordic Conference on Human-Computer Interaction*. NordiCHI '06. ACM, Oslo, Norway, 145–154.

MEIJER, P. B. 1992. An experimental system for auditory image representations. *IEEE Transactions on Biomedical Engineering 39,* 2, 112–121.

MICHAUD, L. AND MCCOY, K. 2006. Capturing the evolution of gramatical knowledge in a CALL system for deaf learners of english. *International Journal of Artificial Intelligence in Education 16,* 1, 65–97.

NAYAK, A. AND BARNER, K. 2004. Optimal halftoning for tactile imaging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 12*, 216–227.

NEWELL, A., ARNOTT, J., BOOTH, L., BEATTIE, W., BROPHY, B., AND RICKETTS, I. 1992. Effect of the "PAL" word prediction system on the quality and quantity of text generation. *Augmentative and Alternative Communication 8,* 4, 304–311.

NSF COMMITTEE ON EQUAL OPPORTUNITIES IN SCIENCE AND ENGINEERING. 2000. Enhancing the diversity of the science and engineering workforce to sustain america's leadership in the 21st century: Executive summary of the CEOSE report to congress, section 6. http://www.nsf.gov/pubs/2001/ceose2000rpt/start.htm. Date accessed: 3 June 2011.

O'DONNELL, M., MELLISH, C., OBERLANDER, J., AND KNOTT, A. 2001. ILEX: an architecture for a dynamic hypertext generation system. *Natural Language Engineering 7,* 3, 225–250.

PINKER, S. 1990. A theory of graph comprehension. In *Artificial Intelligence and the Future of Testing*, R. Freedle, Ed. Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey, 73–126.

Porter. Porter stemmer. http://snowball.tartarus.org/algorithms/porter/stemmer.html.

RAMLOLL, R., YU, W., BREWSTER, S., RIEDEL, B., BURTON, M., AND DIMIGEN, G. 2000. Constructing sonified haptic line graphs for the blind student: first steps. In *Proceedings of the Fourth International ACM Conference on Assistive Technologies*. ASSETS 2000. ACM, Arlington, Virginia, 17–25.

RODGERS, J. L. AND NICEWANDER, W. A. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician 42,* 1, 59–66.

ROTH, P., KAMEL, H., PETRUCCI, L., AND PUN, T. 2002. A comparison of three nonvisual methods for presenting scientific graphs. *Journal of Visual Impairment & Blindness 96,* 6, 420–428.

SCHWARTZ, E. 2007. A browser extension for providing visually impaired users access to the content of bar charts on the web. *Honors Thesis - Millersville University*.

SINHA, R. AND MIHALCEA, R. 2007. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *Proceedings of the First IEEE International Conference on Semantic Computing*. ICSC 2007. IEEE Computer Society, Irvine, California, 363–369.

TAN, P.-N., STEINBACH, M., AND KUMAR, V. 2005. *Introduction to Data Mining*. Addison Wesley.

TODMAN, J. AND ALM, N. 2003. Modelling conversational pragmatics in communication aids. *Journal of Pragmatics 35,* 4, 523–538.

TreTagger. Treetagger. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger.

TRNKA, K., MCCAW, J., YARRINGTON, D., MCCOY, K. F., AND PENNINGTON, C. 2009. User interaction with word prediction: The effects of prediction quality. *ACM Transactions on Accessible Computing 1*, 17:1–17:34.

TUFTE, E. 1983. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut.

WALKER, B. N. AND NEES, M. A. 2005. An agenda for research and development of multimodal graphs. In *Proceedings of the 11th Meeting of the ICAD: International Conference on Auditory Display*. ICAD05. ICAD, Limerick, Ireland, 428–432.

WALKER, M., RAMBOW, O., AND ROGATI, M. 2002. Training a sentence planner for spoken dialogue using boosting. *Computer Speech and Language: Special Issue on Spoken Language Generation 16,* 3, 409–434.

WALLER, A., BROUMLEY, L., AND NEWELL, A. 1992. Incorporating conversational narratives in an AAC device. Presented at ISAAC-92. Abstract appears in *Augmentative and Alternative Communication* 8.

WALLER, A., O'MARA, D., TAIT, L., BOOTH, L., BROPHY-ARNOTT, B., AND HOOD, H. 2001. Using written stories to support the use of narrative in conversational interactions: Case study. *Augmentative and Alternative Communication 17,* 4, 221–232.

WAY, T. P. AND BARNER, K. E. 1997a. Automatic visual to tactile translation–part I: Human factors, access methods and image manipulation. *IEEE Transactions on Rehabilitation Engineering 5,* 1, 81–94.

WAY, T. P. AND BARNER, K. E. 1997b. Automatic visual to tactile translation–part II: Evaluation of the TACTile image creation system. *IEEE Transactions on Rehabilitation Engineering 5,* 1, 95–105.

WICKENS, C. D. AND CARSWELL, C. M. 1995. The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors 37,* 3, 473–494.

WITTEN, I. AND EIBE FRANK, M. A. H. 2011. *Practical Machine Learning: Tools and Techniques, 3rd edition*. Morgan Kaufmann.

WU, P. 2012. Recognizing the intended message of line graphs: Methodology and applications. Ph.D. thesis, University of Delaware.

WU, P., CARBERRY, S., AND ELZER, S. 2010. Segmenting line graphs into trends. In *Proceedings of the Twelvth International Conference on Artificial Intelligence*. 697–703.

WU, P., CARBERRY, S., ELZER, S., AND CHESTER, D. 2010. Recognizing the intended message of line graphs. In *Proceedings of the Sixth International Conference on the Theory and Application of Diagrams*. Diagrams 2010. Springer-Verlag, Portland, Oregon, 220–234.

YU, W. AND BREWSTER, S. 2002. Comparing two haptic interfaces for multimodal graph rendering. In *Proceedings of the 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*. HAPTICS '02. IEEE, Orlando, 3–9.

YU, W. AND BREWSTER, S. 2003. Evaluation of multimodal graphs for blind people. *Universal Access in the Information Society 2,* 2, 105–124.

YU, W., KANGAS, K., AND BREWSTER, S. 2003. Web-based haptic applications for blind people to create virtual graphs. In *Proceedings of the 11th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*. HAPTICS '03. IEEE, Los Angeles, 318–325.

ZACKS, J. AND TVERSKY, B. 1999. Bars and lines: A study of graphic communication. *Memory and Cognition 27,* 6, 1073–1079.

ZHAI, C. 2008. *Statistical Language Models for Information Retrieval*. Synthesis Lectures on Human Language Technologies Series, vol. 1. Morgan & Claypool, San Rafael, California.