CONTENTS

Issues in Realizing the Overall Message of a Bar Chart

Seniz Demir[*], Sandra Carberry[*] & Stephanie Elzer[**]

[*] *University of Delaware*
[**] *Millersville University*

Abstract

Information graphics, such as bar charts and line graphs, that appear in popular media generally have a message that they are intended to convey. We have developed a Bayesian network that recognizes the overall message of a bar chart and produces the logical representation of that message. However, the realization of a logical representation in natural language requires constructing referents for certain graphical elements. This chapter presents our solution to one aspect of this problem: identifying an appropriate referent for the dependent axis. An evaluation study validates our methodology and shows that it is much better than several baseline strategies.

1 Introduction

Information graphics, such as bar charts and line graphs, have been widely used to effectively depict quantitative data and the relations among them. Although in some cases graphics are stand-alone and constitute the entire document, they are usually part of a multimodal document (both text and graphics) where they play an important role in achieving the discourse purpose of the document. Clark (1996) contends that language is not just text and utterances, but instead includes any deliberate signal that is intended to convey a message. Most graphics in popular media contain such signals and are intended to convey a message that is partially enabled by the reader's recognition of these signals. Thus, under Clark's definition, information graphics are a form of language.

It is quite often the case that little or none of a graphic's message in a multimodal document is captured by the article's accompanying text (Carberry et al. 2006). Thus information graphics cannot be ignored. One might suggest relying on a graphic's caption. Unfortunately, graphic captions are often very general and of limited utility in identifying the graphic's message (Corio & Lapalme 1999). For example, the captions on the graphics in Figure 2-a capture little of what the graphics convey. Consequently, information graphics in a multimodal document must be analyzed and comprehended.

Although more and more documents are becoming available electronically, these resources are generally provided in a single format and are not readily accessible to everyone. For example, individuals with sight impairments can access the text in such documents via screen reader programs but they have difficulty when it comes to the graphics. Graphic designers have been encouraged to provide alt text with the graphics but this is seldom done. Researchers have attempted to convey graphics via alternative modalities, such as touch (Ina 1996), or sound (Meijer 1992), or even textual descriptions of the data presented (Ferres et al. 2007, Yu et al. 2007). These approaches have serious limitations such as requiring expensive equipment or requiring that the user develop a mental map of the graphic, something that is difficult for users who are congenitally blind (Kennel 1996).

We are developing an interactive natural language system called SIGHT (Elzer et al. 2007) which is significantly different from previous approaches. SIGHT has the goal of providing the user with the message and knowledge that one would gain from viewing the graphic, rather than providing alternative access to what the graphic looks like. It is envisioned to first provide the user with a brief textual summary of the graphic with the inferred overall message as the core content, and then respond to follow-up questions which may request further detail about the graphic. One problem that we encountered in constructing a natural language version of a graphic's overall message is identifying the appropriate referent for the dependent axis. This chapter presents our implemented methodology for addressing this problem. Although our work has thus far focused on bar charts, we believe that our methodology is extensible to other kinds of graphics.

2 The overall message of an information graphic

The current SIGHT implementation contains a module that uses Bayesian reasoning to hypothesize the overall message of a graphic (Elzer et al. 2007). This module takes as input an XML representation of the graphic (produced by a visual extraction module) that specifies the components of the graphic such as the number of bars and their heights. It exploits a variety of communicative signals present in the graphic (e.g., the salience of entities in the graphic and the presence of suggestive verbs, such as *rising*, in the caption) to infer the overall message of the graphic and produces a logical representation of that message. The logical representation falls into one of twelve message categories that have been identified as the kinds of messages that can be conveyed by a bar chart, such as conveying a change in trend **(Change_Trend)** and conveying the rank of an entity **(Get_Rank)**. Since the overall message of the graphic forms the core of the initial summary in SIGHT, its logical representation must be translated into natural

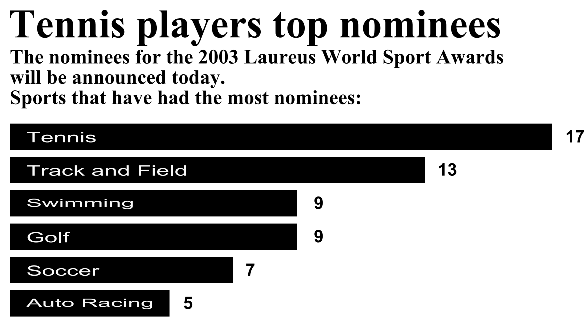

Fig. 1: *Graphic from USA Today*

language. For example, the logical representation of the overall message for the graphic in Figure 1 (**Maximum (First_Bar)**) might be realized in natural language as "**The number of Laureus World Sports's nominees** is highest for Tennis among **the sports** listed: *Track and Field, Swimming, Golf, Soccer, Auto Racing, and Tennis.*"

We observed that the possible realizations of all of these message categories require **a referent for the dependent axis** (e.g., the number of Laureus World Sports's nominees). The referent for the dependent axis is not part of the logical representation and is often not explicitly given in the graphic, as seen in Figure 1. Therefore, the appropriate referent must be extracted from the text of the graphic. In the rest of this chapter, we will describe our approach for constructing this referent.

3 Measurement axis descriptor

3.1 *Corpus analysis*

Information graphics often do not label the dependent axis with a full descriptor of what is being measured in the graphic, but a common requirement for realizing the graphic's overall message is the identification of an appropriate referent for the dependent axis. We will call this referent the **measurement axis descriptor**. We undertook a corpus analysis in order to identify where the measurement axis descriptor appears in a graphic and to motivate heuristics for extracting it. We collected 82 groups of graphics, along with their articles, from 11 different magazines (such as Newsweek and Business Week) and newspapers. We selected at least one bar chart from each group,[1] and our corpus contained a total of 107 simple bar charts.

[1] In cases where a group contains more than one graphic, only simple bar charts in the group were collected.

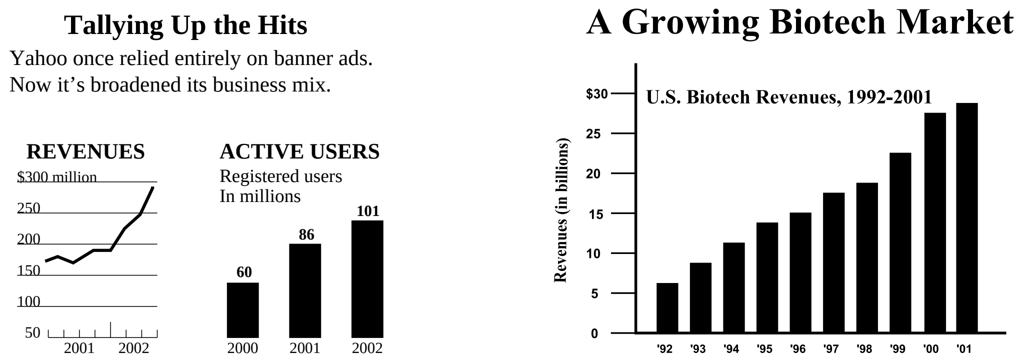

Fig. 2: *(a)Composite graph from Newsweek (b)Graph from Business Week*

Graphic designers generally use text within and around the graphic to present information related to the graphic. We observed that graphics contain a set of component texts that are visually distinguished from one another (e.g., by placement or blank lines), which we refer to as **text levels**. We observed seven text levels, but every level does not have to appear in every graphic. Overall_Caption and Overall_Description apply to composite graphs that contain more than one graphic such as Figure 2-a. In composite graphs, Overall_Caption is the text that appears at the top of the overall group and serves as a caption for the whole set (such as "Tallying Up the Hits" in Figure 2-a). Often there is another text component following the Overall_Caption and distinguished from it by a line or a change in font. This text often elaborates on the set of graphics in the composite graph. We refer to such text as the Overall_Description (such as "Yahoo once relied entirely on banner ads. Now it's broadened its business mix" in Figure 2-a). Caption and Description serve the same roles for an individual graphic. For example, the Caption for the bar chart in Figure 2-a is *"Active Users"* and the Description is *"Registered users In millions"*. The Caption of Figure 2-b is *"A Growing Biotech Market"* but this graphic does not have a Description. There is sometimes a label on the dependent axis itself and we refer to it as Dependent_Axis_Label (such as "Revenues (in billions)" in Figure 2-b). In addition to the text levels described so far, we have observed that there is often a text component residing within the borders of the graphic which we refer to as Text_In_Graphic (such as "U.S. Biotech Revenues, 1992-2001" in Figure 2-b). Finally, Text_Under_Graphic is the text under a graphic which usually starts with a marker symbol (such as *) and is essentially a footnote. Table 1 lists the various text levels, along with how often they appeared in the graphics in our corpus.

Two annotators analyzed each of the 107 graphics in our corpus and determined how they would render the identified message in natural language, with particular attention given to the ideal realization of the measurement axis descriptor. After the descriptors were identified, an analysis was done to see how they were constructed. In 55.1% of the graphics, the ideal measurement axis descriptor appeared as a unit in a single text level, but in 36.5% of these instances, the text level contained additional information. In 44.9% of the graphics, pieces of the measurement axis descriptor had to be extracted from more than one text level and melded together. In these instances, the ideal measurement axis descriptor can be viewed as consisting of a **core** or basic noun phrase from one text level that must be augmented with text from another level (or in some cases, from text in the accompanying article). For example, for the bar chart in Figure 2-a, the information needed to identify the pieces of the descriptor "The number of Yahoo's registered users" must be extracted from two text levels: "Yahoo" from Overall_Description and "registered users" (the core) from Description.

Text level	**Frequency**
Overall_Caption	31.8%
Overall_Description	17.8%
Caption	99.0%
Description	54.2%
Text_In_Graphic	39.3%
Dependent_Axis_Label	18.7%
Text_Under_Graphic	7.5%

Table 1: *Text levels in bar charts*

With the exception of Text_Under_Graphic, the ordering of text levels in Table 1 forms a hierarchy of textual components, with Overall_Caption and Dependent_Axis_Label respectively at the top and bottom. We observed that the core of the measurement axis descriptor generally appears in the lowest text level present in the graphic. This observation is not surprising since text levels lower in the hierarchy are more specific to the graphic's content and thus more likely to contain the core of the ideal descriptor.

During the corpus analysis, we observed three ways in which the core was augmented to produce the ideal measurement axis descriptor:

- **Expansion of the noun phrase:** Nouns in the core of the descriptor were replaced with a noun phrase which has the same noun as its head. For example in Figure 2-b, the core of the descriptor is "Revenues". This noun is reasonable enough to be the core but it should be replaced with "U.S. Biotech Revenues" in order to be complete.
- **Specialization of the noun phrase:** The core was augmented with a proper noun which specialized the descriptor to a specific entity. Figure 2-a shows a composite graph where individual graphics present

different attributes of the same entity (*"Yahoo"*). The ideal measurement axis descriptor for the bar chart (*"Yahoo's registered users"*) consists of the core *"registered users"* augmented with the proper noun "Yahoo" that appears in the Overall_Description.

- **Addition of detail:** Text_Under_Graphic typically serves as a footnote to give specialized detail about the graphic which is not as important as the information given in other text levels. If the Text_Under_Graphic begins with a footnote marker (e.g., *), and the core is followed by the same marker, then Text_Under_Graphic adds detail to the core.

3.2 *Methodology*

Our implemented methodology for constructing a measurement axis descriptor is based on the insights gained from our corpus analysis. First, preprocessing extracts the scale and unit indicators from the text levels or from labels on the dependent axis. For example, the label $90 would indicate that dollar is the unit of measurement. Next heuristics are used to construct the core of the descriptor. Three kinds of augmentation rules are then applied to the core to produce the measurement axis descriptor. Finally, if the descriptor does not already contain the unit of measurement (such as *percent*), the phrase indicating the unit is appended to the front.

3.2.1 *Heuristics*

We developed 9 heuristics for identifying the core of the measurement axis descriptor. The application of the heuristics gives preference to text levels that are lower in the hierarchy, and the heuristics themselves take into account the presence of cue phrases, special characters, and the presence and position of noun phrases in a text level. The heuristics are designed to be dependent on the parses of the text levels. We apply the first two specialized heuristics to Dependent_Axis_Label and Text_In_Graphic respectively. The remaining heuristics are then applied, in order, to a text level, starting with the Description; if a core is not identified at one level, the heuristics are applied, in order, to the next higher level in the hierarchy. All sentences in each text level are examined starting with the last sentence.

- **Heuristic-1:** If the Dependent_Axis_Label consists of a single noun phrase that is not a scale or unit indicator, that noun phrase is the core of the measurement axis descriptor.
- **Heuristic-2:** If Text_In_Graphic consists of a noun phrase, then that phrase is the core; otherwise, if Text_In_Graphic is a sentence, the noun phrase that is the subject of the sentence is the core.

- **Heuristic-3:** If the current sentence at the text level begins with "Here is" or "Here are"(e.g., "Here's a look at the index of leading economic indicators"), the core is the object of the subsequent prepositional phrase (if any); otherwise the core is the noun phrase following "Here is" or "Here are".
- **Heuristic-4:** If the current text level consists of a wh-phrase followed by a colon (:) or a question mark (?), that wh-phrase is the core.
- **Heuristic-5:** If a fragment at the text level consists of a noun phrase followed by a colon (:), and the noun phrase is not a proper noun, that noun phrase is the core.
- **Heuristic-6:** If a fragment at the text level consists of a noun phrase which is not solely a proper noun, that noun phrase is the core.
- **Heuristic-7:** If the current sentence at the text level is followed by a colon (:), the core is the noun phrase preceding the verb phrase in that sentence.
- **Heuristic-8:** The core is the noun phrase preceding the verb phrase in the current sentence at the text level.
- **Heuristic-9:** If the fragment at the text level consists of a proper noun, the possessive form of the proper noun concatenated with the first noun phrase extracted from the closest higher level forms the core. If no such noun phrase is found, only that proper noun forms the core.

In some graphics, what is extracted as the core is a complex noun phrase whose head matches the ontological category of the bar labels. Clearly, this category refers to the independent axis and is not what the dependent axis is conveying. In such cases, the head is either modified by a "relative clause" or followed by a phrase beginning with "with". This phrase or relative clause tends to define the aspect of the bars that is being measured by the dependent axis. Therefore, the nouns and subsequent prepositional phrases in the modifier are instead collected as the core. For example our heuristics would initially extract "Sports that have had the most nominees" as the core in Figure 1; since sports is the category of the labels, "nominees" becomes the core.

3.2.2 *Augmentation rules*

We have defined three augmentation rules that correspond to the three kinds of augmentation observed during the corpus analysis. If none of the augmentation rules is applicable in a graphic, the core forms the full descriptor.

- **Expansion of the noun phrase:** To expand a noun phrase, examine text levels higher in the hierarchy than the text level from which the core was extracted; if a noun phrase appears with the same head noun as a noun in the core, then the noun in the core is replaced with the larger noun phrase.

- **Specialization of the noun phrase:** To specialize a noun phrase, determine whether i) there is only one proper noun at all text levels higher in the hierarchy than the level from which the core was extracted, or ii) there is only one proper noun in the Overall_Caption or the Caption. If one of these two criteria are satisfied and the proper noun is not a bar label in the graphic, then the possessive form of that proper noun is appended to the front of the core.
- **Addition of detail:** To add detail to a core, determine whether the core was followed by a footnote marker in the text level from which it was extracted; if so, Text_Under_Graphic that is preceded by the same marker is appended to the core as a bracketed expression.

4 Examples of referent identification

For the graphic in Figure 1, Heuristic-5 initially identifies the noun phrase *"sports that have had the most nominees"* as the core. However, its head noun *"sports"* matches the ontological category of the bar labels; consequently, the noun *"nominees"* in the relative clause modifying *"sports"* becomes the core. The augmentation rule for specialization finds that *"Laureus World Sports"* is the only proper noun in the text levels and constructs *"Laureus World Sports's nominees"*. After adding a pre-fragment representing the unit of measurement, the referent for the dependent axis becomes *"The number of Laureus World Sports's nominees"*.

For the graphic in Figure 2-b, Heuristic-1 identifies *"Revenues"* in Dependent_Axis_Label as the core. Since the core and the Text_In_Graphic, *"U.S. Biotech Revenues"*, have the same head noun, the augmentation rule for expansion produces *"U.S. Biotech Revenues"* as the augmented core. After adding a pre-fragment, the referent for the dependent axis becomes *"The dollar value of U.S. Biotech Revenues"*.

5 Evaluation of the implemented system

We constructed a test corpus consisting of 205 randomly selected bar charts from 21 different newspapers and magazines; only six of these sources were also used to collect the bar charts for the corpus study. For each graphic, we used our system to identify the measurement axis descriptor, and the resultant descriptor was rated by two evaluators. The evaluators each assigned a rating from 1 to 5 (with 5 being the best) to the system's output; if the evaluators differed in their ratings, then the lowest rating was recorded. For comparison, three baselines were computed, consisting of evaluations of the

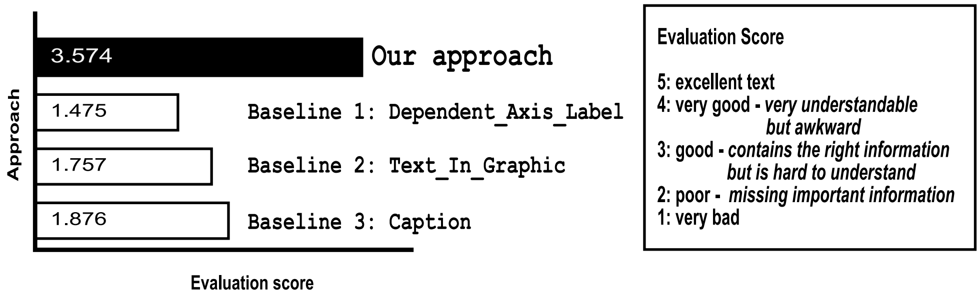

Fig. 3: *Evaluation of the resultant measurement axis descriptor*

text that would be produced using each of the following three text levels as the measurement axis descriptor: Dependent_Axis_Label, Text_In_Graphic, and Caption. For the baselines, if the evaluators differed in their rating of the resultant output, the higher rating was recorded, thereby biasing our evaluation toward better scores for the baselines (in contrast with the scores for our system's output). The results of the evaluation, which are presented in Figure 3, show that our system produces measurement axis descriptors that rate midway between good and very good. It is particularly noteworthy that our methodology performs much better than any of the baseline approaches. However, further work is still needed to improve our results, such as resolving pronominal references within the text.

We computed how frequently each heuristic was applied in the corpus used for the empirical study and in the test corpus. As shown in Table 2, each heuristic was applied at least once in both corpuses. Heuristic_2 and Heuristic_6 were the most frequently applied heuristics in both corpuses. We observed that the number and the kinds of heuristics that were applied to the graphics collected from the same media varied for the two corpuses. For example, Heuristic_6 and Heuristic_8 were the only heuristics applied to the graphics collected from the News Journal in the empirical study corpus. But in the test corpus, five different heuristics were applied to the graphics collected from the same newspaper. 41 graphics in the test corpus were taken from newspapers and magazines that were not used in the empirical study. Thus our success on the test corpus suggests that our methodology is not limited to the specific media on which the empirical study was based.

Corpus	**H_1**	**H_2**	**H_3**	**H_4**	**H_5**	**H_6**	**H_7**	**H_8**	**H_9**
Corpus Study	1	36	4	11	11	32	2	7	3
Test Corpus	2	69	1	5	32	76	7	12	1

Table 2: *Frequency of use of the heuristics*

6 Conclusion

This chapter has focused on an issue that needs to be addressed in realizing the inferred overall message of a bar chart: constructing a referent for the dependent axis. Our methodology for constructing an ideal measurement axis descriptor uses heuristics to first identify a core descriptor and then augments that core. We presented our corpus study which provides the insights that were used in developing the realization methodology. Evaluation of our approach shows that our methodology generally produces reasonable text, and that it performs far better than any of three baseline approaches.

Acknowledgements. This material is based upon work supported by the National Science Foundation under Grant No. IIS-0534948.

REFERENCES

Carberry, Sandra , Stephanie Elzer & Seniz Demir. 2006. "Information Graphics: An Untapped Resource for Digital Libraries". *Proceedings of the 29th Annual Conference on Research and Development on Information Retrieval*, 581–588. Seattle, Washington.

Clark, Herbert H. 1996. *Using Language*. Cambridge: Cambridge University Press.

Corio, Marc & Guy Lapalme. 1999. "Generation of Texts for Information Graphics". *Proceedings of the 7th European Workshop on Natural Language Generation*, 49-58. Maine, France.

Elzer, Stephanie, Edward Schwartz, Sandra Carberry, Daniel Chester, Seniz Demir & Peng Wu. 2007. "A Browser Extension for Providing Visually Impaired Users Access to the Content of Bar Charts on the Web". *Proceedings of the 3rd International Conference on Web Information Systems and Technologies*, 59-66. Barcelona, Spain.

Ferres, Leo , Petro Verkhogliad, Gitte Lindgaard, Louis Boucher, Antoine Chretien & Martin Lachance. 2007. "Improving Accessibility to Statistical Graphs: The iGraph-Lite System". *Proceedings of the 9th International Conference on Computers and Accessibility*, 67-74. Tempe, Arizona.

Ina, Satoshi. 1996. "Computer Graphics for the Blind". *ACM SIGCAPH Computers and the Physically Handicapped* 55.16-23.

Kennel, Andrea R. 1996. "Audiograf: A Diagram-Reader for the Blind". *Proceedings of the 2nd International Conference on Computers and Accessibility*, 51-56. Vancouver, British Columbia.

Meijer, Peter B. 1992. "An Experimental System for Auditory Image Representations". *IEEE Transactions on Biomedical Engineering* 39:2.112-121.

Yu, Jin, Ehud Reiter, Jim Hunter & Chris Mellish. 2007. "Choosing the Content of Textual Summaries of Large Time-series Data Sets". *Natural Language Engineering* 13:1.25-49.

Index of Subjects and Terms