In: Text Summarization Branches Out (workshop) ACL, Barcelona, 25-26 July 2004

Handling figures in document summarization

Robert P. Futrelle
Biological Knowledge Laboratory
College of Computer & Information Science WVH202
Northeastern University
Boston, MA 02115, USA
futrelle@ccs.neu.edu

Abstract

Some document genres contain a large number of figures. This position paper outlines approaches to diagram summarization that can augment the many well-developed techniques of text summarization. We discuss figures as surrogates for entire documents, thumbnails, extraction, the relations between text and figures as well as how automation might be achieved. The focus is on diagrams (line drawings) because they allow parsing techniques to be used, in contrast to the difficulties of general image understanding. We describe the advances in raster image vectorization and parsing needed to produce corpora for diagram summarization.

1 Introduction

Many documents contain figures, both images and line drawings (diagrams). The Biology literature, the focus of our group's work, is replete with figures. A surprising 50% of the typical Biology paper is figure-related (see Appendix A). A million or so Biology papers are published each year, most with abstracts. But given their high figural content, work on *diagram summarization* could also be quite useful. This is a position paper that explores this topic, outlining a variety of approaches to the construction of automated diagram summarization systems. System building and use awaits the creation of the requisite corpora, as explained in the next section. This paper builds on our earlier, more lengthy work, extending it in various ways (Futrelle, 1999).

2 The current state of the Diagrams field

Automated text summarization has at its disposal, electronic documents that allow the use of all the techniques of computational linguistics. Diagrams in documents are in a more primitive state. The overwhelming majority of diagrams available in the electronic forms of the research literature today are in raster format. What is needed are diagrams in *vector format* in which an object such as a line is represented not by pixels, but as a line *object* defined by its endpoints and width. We found only 52 pages containing vector-based diagrams in a collection of 20,000 recent Biology papers in PDF format (Futrelle, Shao, Cieslik, & Grimes, 2003). *Vectorization* converts raster diagrams to vector format, much as OCR converts rasters to characters. But the resulting vectorized diagram is an unordered collection of objects in two dimensional space. An additional analysis step of *parsing* is required. Our system for parsing diagrams (Futrelle & Nikolakis, 1995) produces descriptions for a data graph, for example, by discovering structures such as scale lines and sets of data points.

There appear to be no non-proprietary vectorization systems that are up to the task of vectorizing the diagrams from the scientific literature, so our group is currently focused on developing a system for this in Java. We are also redeveloping our parsing system in Java. Until this work is completed, there will be few diagrams available for the application of diagram summarization techniques. This notwithstanding, diagram summari-

zation is an interesting and ultimately important task, which is why we are discussing it here. This work is part of our laboratory's long-term effort to characterize the conceptual content of the Biology literature, including the text and figural content.

3 Figures as surrogate documents

Some time ago, when Lesk asked chemists what two parts of Chemistry papers would be most informative, they said they would like to know the names of the authors and to see the figures (Michael Lesk, personal communication).

Recently, journals are beginning to implement approaches in this spirit. The *Journal of Proteome Research* lists in the table of contents, in both the print and online editions, an entry for each paper that includes the title, authors, abstract and one uncaptioned figure from the paper, typically in color. *Science* and *Nature* also include some figures in their contents pages. The new open-access journal, *PLoS Biology*, offers five "Views" of a paper: HTML, Tables, Figures, Print PDF and Screen PDF. The Figures View is an HTML slide show of the figures, each including a large version of the figure, the caption and the article citation.

Figure Views represent a new and important type of summary of entire articles, allowing the rapid browsing that such visual displays provide. One can imagine that authors will adapt to this new mode, packing the major content of their papers into the figures and captions, reducing the need to read the full text.

4 Thumbnail images are summaries

Thumbnails are images that have been reduced in size and/or cropped to a smaller size. Shrinking an entire image so that it acts as a summary is an analog operation that has no parallel in text. For some images, shrinking them too much can produce an illegible result, a practice that has been roundly criticized (item 4 in Nielsen, 2003); cropped images may be useful in such cases.

Figure 1. A full-scale *analog extract* (1% of the original) of the "classic" London Underground map. This is an *informative summary* with respect to the map style, but is only *indicative* of the full map.

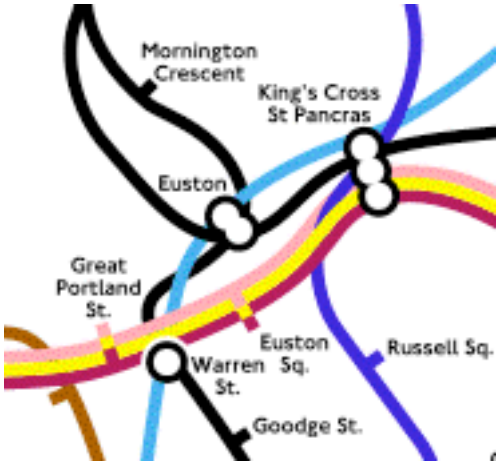

Figure 2. The same type of summary extract, except that it is taken from a geographically correct view of the same section of the Underground as shown in Fig. 1.

An example of cropping two very large images resulting in informative thumbnails appears in the Figure Gallery item on our site, `http://diagrams.org/fig_pages/f00022.htm` The thumbnails are reproduced here in Figures 1 and 2.

5 Extraction for summarization

One of the most important techniques used in text summarization, is extraction, typically the extraction of carefully chosen whole sentences. A

similar approach can be used for diagram summarization, but some thought needs to be given to what the sentence-like elements in diagrams might be. It is not difficult to give examples of diagram extraction, but automating it is difficult.

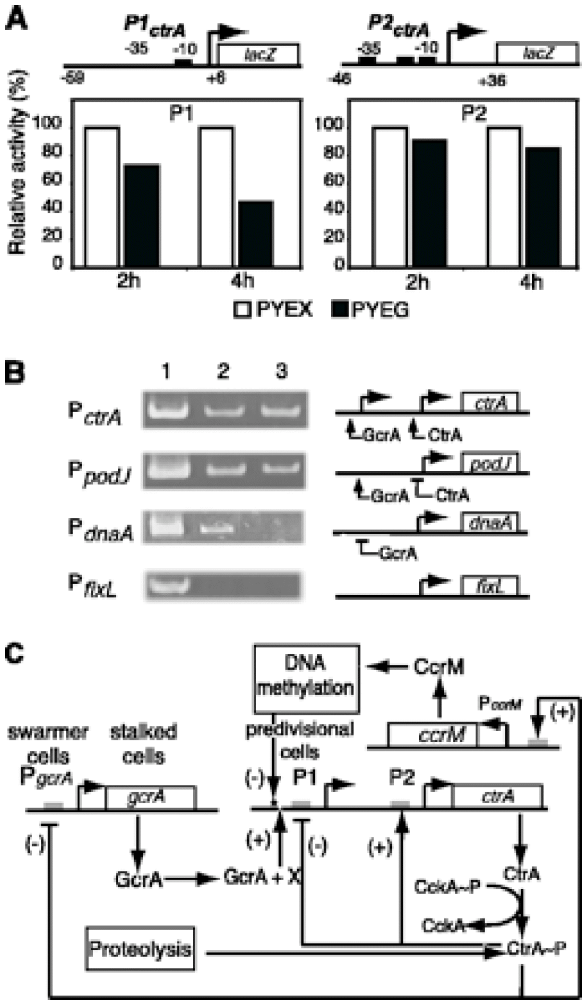

Figure 3. A typical diagram that allows summarization by extraction. From (Holtzendorff, 2004). In this case, retention of one of the two bar graphs in A, one of the four rows in B and all of C would result in a modest, indicative summary of the three-part figure. The keys at the bottom of part A would have to be retained.

Fig. 3, from an issue of *Science*, is typical of diagrams that appear in the Biology research literature. The extraction suggested in our caption picks one item from each of two sets of similar items to produce and indicative summary.

6 Diagram-related text

It might be argued that the most salient content of documents with figures can be found in the text; that the figures are redundant, merely "illustrative". This is often not the case. There are queries to documents that cannot be answered based on the content of the text or diagrams considered separately (Futrelle & Rumshisky, 2001). In Biology it is not unusual for a caption to explain only the methods used to produce the data shown.

The independent contribution of diagram content to a paper is often signaled by *cue phrases*. In referring to data graphs, phrases such as "shows a significant difference" or "are similar" or "a pronounced effect" require that the reader examine the data shown in the figure in order to understand what the phrases refer to.

Fig. 4 (Nijhout, 2003) appeared in the popular scientific journal, *American Scientist,* and is more carefully explained than most. The Fig. 4 caption text illustrates some limitations of captions. For example, the phrase, "The possible combinations" does not spell out what combinations are possible or are illustrated. The reader must study the figure to discover that there are in fact three distinct combinations.

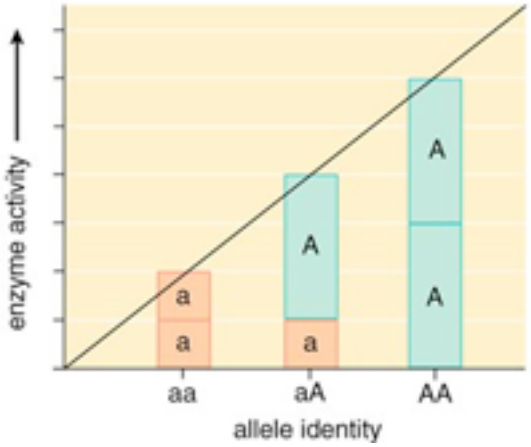

Figure 4. The original caption for this figure, with bolding added, was: " Enzyme activity is a function of allele identity. In this example, the allele **A** encodes an enzyme that has three times greater activity than the enzyme encoded by allele **a**. The possible combinations of **A** and **a** in an individual yield a wide range of overall activity levels."

The references to ***A*** and ***a*** in Fig. 4, are *deictic references*, pointing to objects visible in the context, in the figure. In ordinary conversation, such a reference would point to some physical object in the view of the listener.

A summarization of Figure 4 should include the entire diagram. The last sentence of the caption would be a suitable summary of the caption.

The non-caption text and the text within figures play important roles and need to be taken into account in any attempt to produce a summary. Space precludes further discussion of these.

7 Prospects for automation

Some degree of summarization might be possible based entirely on the classes of the diagrams or subdiagrams in a paper. We have been able to locate subdiagrams in vector-based diagrams in PDFs and successfully classify them using SVMs (Futrelle, Shao, Cieslik, & Grimes, 2003).

But any more detailed summarization decisions would require parsed representations of the diagrams. For example, our parser can discover and analyze the two bar charts in Fig. 3, allowing a system to extract only one of them, though without any knowledge as to which is the most salient. The parser can also locate keys, such as the ones in Fig. 3, so they can be extracted also. Standard strategies from text summarization, such as extracting the diagrams most often referred to, diagrams appearing near the beginning and end of a paper, etc., are all possible. Clearly, automation of diagram summarization presents a new set of challenges and is no easier than text summarization.

Large scale evaluation of diagram summarization will offer its own challenges, cf. text summarization evaluation (Radev et al., 2003).

8 Related work

Automated text summarization has advanced substantially in the last decade. See for example, the major collection of papers, (Mani & Maybury, 1999) and the special journal issue (Radev, Hovy, & McKeown, 2002). Reviews include (Hovy, 2002; Marcu, 2003). A recent useful monograph is (Mani, 2001). Another recent work is (Barzilay, 2003), focused on multidocument summarization and going beyond sentence extraction to consider phrases.

Paradoxically, work on the summarization of scientific articles is inhibited by the fact that virtually all scientific articles have abstracts as a standard component. But there are other tasks such as developing user-tailored summaries (Teufel & Moens, 2002).

The generation of coordinated explanations involving text and graphics offers insight into the relations between them (Feiner & McKeown, 1990). This task involves dealing with the internal structure of diagrams, as do problems of image retrieval, which can be aided by developing ontology-based descriptions of the images (Hyvönen, Styrman, & Saarela, 2002).

Diagrams form a part of a coordinated discourse, so that diagram summarization can profit from the work done on text summarization that focuses on discourse structure. Examples of discourse-related approaches include (Boguraev & Neff, 2000; Marcu, 1997a, 1997b; Teufel & Moens, 2002).

9 Conclusions

Document summarization including diagrams seems both possible and desirable. Work in this area is waiting on the development of a corpus of parsed object-based diagrams. The vectorization and parsing systems required are under development.

Acknowledgement. This material is based upon work supported by the National Science Foundation under Grants No. DBI-0211047 and IIS-9978004 and the Northeastern University Institute for Complex Scientific Software, `http://www.icss.neu.edu/`.

References

Barzilay, R. (2003). Information Fusion for Mutlidocument Summarization: Paraphrasing and Generation. Unpublished PhD, Columbia University.

Boguraev, B. K., & Neff, M. S. (2000). Discourse Segmentation in Aid of Document Summarization. In Proceedings of the 33rd Hawaii International Conference on System Sciences (pp. 10).

Feiner, S. K., & McKeown, K. R. (1990). Coordinating Text and Graphics in Explanation Generation. In AAAI 90 (pp. 442-449).

Futrelle, R. P., & Nikolakis, N. (1995). Efficient Analysis of Complex Diagrams using Constraint-Based Parsing. In ICDAR-95 (Intl. Conf. on Document Analysis & Recognition) (pp. 782-790). Montreal, Canada.

Futrelle, R. P. (1999). Summarization of Diagrams in Documents. In I. Mani & M. Maybury (Eds.), Advances in Automated Text Summarization (pp. 403-421). Cambridge, MA: MIT Press.

Futrelle, R. P., & Rumshisky, A. (2001). Discourse Structure of Text-Graphics Documents. In 1st International Symposium on Smart Graphics. Hawthorne, NY: ACM.

Futrelle, R. P., Shao, M., Cieslik, C., & Grimes, A. E. (2003). Extraction, layout analysis and classification of diagrams in PDF documents. In ICDAR 2003 (Intl. Conf. Document Analysis & Recognition) (pp. 1007-1014). Edinburgh, Scotland: IEEE Computer Society.

Holtzendorff, J., Hung, D., Brende, P., Reisenauer, A., Viollier, P. H., McAdams, H. H., et al. (2004). Oscillating Global Regulators Control the Genetic Circuit Driving a Bacterial Cell Cycle. *Science, 304*, 983-987.

Hyvönen, E., Styrman, A., & Saarela, S. (2002). Ontology-Based Image Retrieval. In Towards the semantic web and web services, Proceedings of XML Finland 2002 Conference (pp. 15-27).

Mani, I. (2001). Automatic Summarization. Amsterdam: John Benjamins.

Mani, I., & Maybury, M. T. (Eds.). (1999). Advances in Automatic Text Summarization. Cambridge, MA: MIT Press.

Marcu, D. (1997a). From discourse structures to text summaries. In I. Mani & M. Maybury (Eds.), Workshop on Intelligent Scalable Text Summarization (pp. 82-88). Madrid, Spain: Assoc. Computational Linguistics.

Marcu, D. (1997b). The rhetorical parsing, summarization, and generation of natural language texts. Unpublished Ph.D., U. Toronto, Toronto.

Nielsen, J. (2003). Jakob Nielsen 's Alertbox, December 22, 2003: Top Ten Web Design Mistakes of 2003. Retrieved May 15, 2004, from http://www.useit.com/alertbox/20031222.html

Nijhout, H. F. (2003). The Importance of Context in Genetics. American Scientist, 91(5), 416-423.

Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. Computational Linguistics, 28(4), 399-408

Radev, D. R., Lam, W., Celebi, A., Teufel, S., Blitzer, J., Liu, D., et al. (2003). Evaluation Challenges in Large-Scale Multi-Document Summarization. In ACL-2003 (pp. 375-382).

Teufel, S., & Moens, M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. Computational Linguistics, 28(4), 409-445.

Appendix A: 50% of the content of Biology papers is figure-related

We arrived at the 50% figure by sampling a variety of recent papers in journals including *Science*, *Nature*, *PNAS (USA)*. The column-inches occupied by figures in the hardcopy or equivalent PDF versions of the papers were measured and compared to the total column-inches, omitting the title, abstract and references. Word counts of the captions and direct running text reference sentences were estimated, e.g., "Fig. 3 shows ...". Then estimates were made of the sentences that indirectly discussed the figures, often the sentences immediately following direct reference sentences and containing anaphoric and definite noun phrase references to the figures, often in deictic form. The total of the figure and figure reference content consistently amounted to about 50% of the papers sampled.