Intelligent Multimedia Interfaces

Mark T. Maybury (editor)
(MITRE Corporation)

Menlo Park, CA: AAAI Press and
Cambridge, MA: The MIT Press, 1993,
vi + 405 pp.
Paperbound, ISBN 0-262-63150-4, $39.95

Reviewed by
Kent Wittenburg
Bellcore

Intelligent Multimedia Interfaces is a follow-up to the AAAI Workshop on Intelligent Multimedia Interfaces held in Anaheim, California, in August 1991. Of interest to computational linguists is the fact that natural language, discourse, and planning research have a much greater presence here than in the previous installment of this line of workshops. A comparable collection of papers, *Intelligent User Interfaces* (Sullivan and Tyler 1991), emerged from a 1988 AAAI/CHI workshop named Architectures for Intelligent Interfaces: Elements and Prototypes. By my informal, ad hoc count, about half of the papers in the earlier collection had something to do with the AI side of computational linguistics, whereas all but perhaps two of the papers in Maybury's volume do.

The work is divided into three sections. Here is the lineup:

Section 1: Automated Presentation Design

S. F. Roth and W. E. Hefley, "Intelligent multimedia presentation systems: Research and principles"

M. T. Maybury, "Planning multimedia explanations using communicative acts"

E. André, W. Finkler, W. Graf, T. Rist, A. Schauder, and W. Wahlster, "WIP: The automatic synthesis of multimodal presentations"

E. André and T. Rist, "The design of illustrated documents as a planning task"

S. K. Feiner and K. R. McKeown, "Automating the generation of coordinated multimedia explanations"

S. K. Feiner, D. J. Litman, K. R. McKeown, and R. J. Passonneau, "Towards coordinated temporal multimedia presentations"

B. A. Goodman, "Multimedia explanations for intelligent training systems"

Section 2: Intelligent Multimedia Interfaces

J. D. Burger and R. J. Marshall, "The application of natural language models to intelligent multimedia"

O. Stock and the ALFRESCO Project Team, "ALFRESCO: Enjoying the combination of natural language processing and hypermedia for information exploration"

S. Abu-Hakima, M. Halasz, and S. Phan, "An approach to hypermedia in diagnostic systems"

D. B. Koons, C. J. Sparrell, and K. R. Thorisson, "Integrating simultaneous input from speech, gaze, and hand gestures"

Section 3: Architectural and Theoretical Issues

Y. Arens, E. H. Hovy, and M. Vossers, "On the knowledge underlying multimedia presentations"

M. Cornell, B. P. Woolf, and D. Suthers, "Using 'live information' in a multimedia framework"

J. Krause, "A multilayered empirical approach to multimodality: Towards mixed solutions of natural language and graphical interfaces"

A. Bonarini, "Modeling issues in multimedia car-driver interaction"

Despite the somewhat misleading organization, most of the papers in this book really are mostly about automatic presentation, whether from a systems or conceptual point of view. The remaining papers don't seem to congeal too easily. There is one about empirical experiments from a human factors perspective (Krause); one about gestures (Koons et al.); three that touch on hypermedia (Stock et al.; Abu-Hakima et al.; Cornell et al.); and one about car-driver interaction (Bonarini). About half of the articles are oriented toward discussing an implemented prototype. The others range from reviews of the state of the art to position papers to discussions of design issues for prototypes of the future.

Lest there be any doubt about the slant taken toward multimedia interfaces most in evidence in this collection, consider these representative quotes:

> Our approach is based on a generalization of a set of rhetorical acts used previously for multisentential text planning. (Maybury, p. 72)
>
> Most semantic and pragmatic relationships which have been proposed for describing the structure of texts can be generalized in such a way that they are also appropriate for describing the structure of pictures and text-picture combinations. (André and Rist, p. 115)
>
> COMET uses three different domain knowledge sources: a static representation of domain objects and actions ..., a diagnostic rule-base, and a detailed geometric knowledge base .... It also maintains a user model and a model of the previous discourse. (Feiner and McKeown, p. 123)
>
> Many of the mechanisms of an intelligent interface can benefit from a rich and explicit representation of knowledge. (Burger and Marshall, p. 177)
>
> The motto is "No ink on the screen," or no presentation without an understanding by the system of the semantics behind that presentation. (Cornell, Woolf, and Suthers, p. 308)

Most of the work in this collection is then cut from the grand AI cloth. Explicit and exhaustive symbolic representation is an important goal for many of these researchers, as is the drive to, in some sense, imitate human abilities through the computer. An oft-unstated, and certainly unproven, assumption is that attempting to imitate human–human communication in human–computer interaction will necessarily improve the latter. Given the title of this book, this emphasis shouldn't be too surprising; the word "intelligent," when applied to systems these days, is often a code word for symbolic, frame-based AI approaches.

But multimedia adds a bit of a twist to the Turing test. Systems exhibiting multimedia (graphics, animation, video, sound, text) on the output side don't much resemble your average Joe on the street. Multimedia presentation may require such things as designing and rendering a pie chart to present certain kinds of data or re-assembling animation clips into a well-designed movie on the fly. Well-trained experts, given the time and the tools, can do this kind of thing, but it is not an innate human ability. And the computer is definitely a useful tool here quite apart from whether it resembles a human emoting device. So can we pin down the meaning of intelligence any further in this context? Here are some behavioral features suggested by the work in this volume that some might use to distinguish the intelligent multimedia systems from the dumb ones:

- Natural language (nonexclusively) in input and/or output, e.g.,
  - augmenting a text-based NL query/response capability with hypermedia or other direct-manipulation graphics (Stock et al.; Burger and Marshall);
  - generating graphics or temporal media along with NL text (André et al.; André and Rist; Feiner and McKeown; Feiner et al.; Arens et al.).
- Coordinated multiple media streams in input and/or output, e.g.,
  - interpretation of gestures with a mouse or data glove, in parallel with spoken command language (Koons et al.);
  - automatic creation and presentation of appropriate textual overlays for animation clips (Goodman);
  - generation of referring expressions dependent upon the presentations in a different media stream (André et al.; André and Rist; Feiner and McKeown).
- Creativity in output generation—that is, moving away from "canned" presentations to more flexible ones, possibly implying:
  - reuse of media objects for multiple purposes, e.g., assembling differing sequences from the same set of video clips depending on the query or goal of the presentation (Goodman);
  - multiple media and presentation types made available for the same semantic information (Cornell et al.);
  - presentation affected by aspects of the type or category of information coordinated with the type or category of media (Roth and Hefley; Arens et al.);
  - content selection and presentation affected by the communicative goals of the presentation or aspects of the user profile or dialog history (Maybury; André et al.; André and Rist; Feiner and McKeown; Burger and Marshall; Bonarini).

I think it's fair to say that this collection is an excellent representation of current work in the area of knowledge-based multimedia, particularly work with a natural language orientation. There are papers from most of the major ongoing research efforts of this kind. The venues include Columbia University; Information Sciences Institute, University of Southern California; IRST in Trento, Italy; MITRE Corporation; Massachusetts Institute of Technology; and DFKI (German AI Institute), Saarbrücken.

For me, highlights begin with the first paper in the volume, Roth and Hefley's "Intelligent multimedia presentation systems: Research and principles." Roth and Hefley produce a very lucid summary and analysis of work to date on automatic presentation, beginning with a conceptual breakdown of the architecture of intelligent multimedia generation systems. They include work in so-called information graphics, along with the more language- and discourse-oriented AI approaches. One of the issues that they give a particularly good account of is data modeling for graphics and multimedia generation. One of the fundamental and largely unanswered questions for many of the more ambitious automatic multimedia generation systems, given their dependence on hand-tuned frame representations for their content data, is how they can be used in conjunction with massive data collections in existing databases. Roth and Hefley's hope is that it may be possible to develop a medium-independent language for describing the characteristics of information that will generalize well across differing data collections.

The next five papers are at the heart of this collection. They include descriptions and discussions of issues surrounding current and planned implementations of prototype systems for automatic synthesis of NL text and graphics. These authors emphasize the role of explicit representation of not only semantic content but also such things as communicative goals, discourse histories, and user models. Maybury writes on aspects of presentation planning as communicative acts, offering a framework for including both linguistic acts and graphical acts. WIP, discussed by André et al. and André and Rist, is an experimental system that synthesizes explanations in natural language and graphics. COMET, a system for generating coordinated explanations involving text and three-dimensional graphics, is covered in the next two papers. Feiner and McKeown give an overview of the architecture of current implementation. Feiner et al. discuss design issues in extending the implementation to deal with coordination of temporal media. WIP and COMET are two of the most ambitious prototypes of AI-based systems for generating coordinated NL text and graphics to date. An impressive number of researchers and graduate students have been involved in the implementations.

Bradley Goodman, in his article titled "Multimedia explanations for intelligent training systems," motivates his work not just from the push of conceptual AI problems but from the pull of applications as well. In contrast to other multimedia explanation research, this work is focused on the re-use of multimedia materials (in particular, video clips) rather than on generation of new material from full underlying semantic representations. Arguing that "re-use is a problem closer to the needs of current training systems than generating multimedia explanations from scratch," he focuses on creative re-use of video clips in explanations for training systems.

The first two papers of Section 2 are closely related to the core papers of Section 1. Burger and Marshall discuss multimedia dialog issues in the system called AIMI, which incorporates natural language text and graphics in a military-map–based domain. The ALFRESCO system is discussed by Stock et al. The innovation here is to embed hyperlinks into the natural language text generated by the system. The application as a whole, which is of the information kiosk variety in the domain of Italian frescos, includes conventional hypermedia functionality along with NL query and generation. The discussion of the synergies between these two quite different technologies (hypermedia and NL query/response) is reminiscent of Cohen et al.'s (1989) landmark work on synergistic use of natural language and direct manipulation graphics.

Aside from the paper "On the knowledge underlying multimedia presentations," by Arens, Hovy, and Vossers, which is an informed and detailed classification of knowledge useful for automatically generating presentations, the remaining articles in the volume are one-of-a-kind. Abu-Hakim, Halasz, and Phan summarize probably the most mature of the implementations appearing in this volume (evidenced among other things by the number of included screen dumps). JETA is a knowledge-based diagnostic system in the domain of jet engine repair. The "intelligence" of this system is not so much in the service of media generation or interpretation in the interface, but in the underlying workings of the diagnostic system. The paper by Koons, Sparrell, and Thorisson, "Integrating simultaneous input from speech, gaze, and hand gestures," is a continuation of this line of work that has been going on at MIT under Robert Bolt's direction for some 15 years. It includes one of the very few pieces of research I have seen that treats the interpretation of gestures as a genuine parsing problem. A rather different application motivates Bonarini's work discussed in "Modeling issues in multimedia car-driver interaction." The simulated copilot of an intelligent vehicle is faced with, for example, directing a driver on a preplanned route through heavy traffic. Various aspects of the driver's mental state and the vehicle's location are modeled and affect the means by which messages are conveyed. The concept of "live information," as explicated by Cornell, Woolf, and Suthers, is reminiscent of the HITS experiment at MCC in the mid-1980s, whose goal was to explicitly represent anything and everything in an interface (Hollan et al. 1991). It is of interest to note that after sporting the button "No Presentation without Representation" for several years, the slogan became "No Representation without Utilization." Various forces had something to do with this turnaround, but I daresay it also reflects the experience of some of the researchers on the project who became discouraged about the overall enterprise (see Wroblewski, McCandless, and Hill 1991). Krause, in fact, sounds one of the few cautionary notes in this volume about the usual AI approach to doing research on multimedia interfaces. In "A multilayered empirical approach to multimodality: Towards mixed solutions of natural language and graphical interfaces," he questions the assumption that striving to imitate human–human communication is the most effective way to improve upon human–computer interaction. In place of this research paradigm, he offers an empirical approach involving the building and testing of prototypes that are informed not by attempting to imitate human–human communication, but by painstaking experiments and evaluation. Clearly, there is a real tension between the goal of improving human–computer interaction and the goal of engineering systems that model human-like abilities.

Religious arguments notwithstanding, the builders of knowledge-based multimedia systems would do well to carry their systems beyond the toy demo stage to at least the point where a community of users can accomplish real tasks with them. Claims about success and effectiveness for HCI sound rather hollow when no evidence is available other than an existence proof that a system runs on a few researcher-generated test cases. As Roth and Hefley point out, there is a wide range of research questions that are simply unanswerable without real users to move things along. But of course there is a tradeoff. So far, all one can say is that most of the systems discussed in this volume weigh in heavily on the proof-of-concept side. Bridging the gap to useful systems has to be an important issue, not only to keep the funders of this sort of research happy, but for the health of the research itself, particularly if effective human–computer interaction is really an end goal.

No review is complete without nitpicking. My first complaint is that the book's organizational scheme seems, let us say, suboptimal. The three sections are, again, (1) "Automated Presentation Design," (2) "Intelligent Multimedia Interfaces," and (3) "Architectural and Theoretical Issues." So Section 2 carries the title of the book. That's odd. The opening article in Section 1, "Intelligent multimedia presentation systems: Research and principles," sounds suspiciously like it ought to go in Section 3. At least one article in Section 3, "On the knowledge underlying multimedia presentations," is specifically about presentation, the topic of Section 1. In the editor's defense, it should be pointed out that the papers are not a particularly easy group to massage into a well-balanced tree given the heavy weighting toward the automatic-presentation side. Maybury does provide solid introductions to the book as a whole and to each section. Another nit is that the book has more than its share of typos and format errors, beginning with the appearance of *multimedia* as *multimedial* in the table of contents. And I don't understand why the volume does not include any information about the authors other than their names: no affiliations, no postal or e-mail addresses. The index, too, seems rather half-hearted.

As I say, these complaints are nits. On the whole, this book is a milestone. It belongs on the shelf of any AI-oriented language researcher interested in branching out to multimedia. I suspect that there are many readers of this journal who are in that category. But those who are looking for human interface ideas that have had some measure of validation (or even much exercise), or for a cookbook of techniques ripe for implementing today's multimedia systems will be disappointed. This largely reflects the state of the field rather than a problem with this particular volume. My hope for the future is that the next collection of papers on advanced multimedia interfaces will include a broader spectrum of work from the computational linguistics community. Already there are results in extending grammars and parsing techniques into multimedia domains. Statistically oriented computational linguistics methods should be able to find a place as well.

References

Cohen, Philip R.; Dalrymple, Mary; Moran, Douglas B.; Pereira, Fernando C. N.; Sullivan, Joseph W.; Gargan, Robert A. Jr; Schlossberg, Jon L.; and Tyler, Sherman W. (1989). "Synergistic use of direct manipulation and natural language." In *Proceedings, CHI '89 (Conference on Human Factors in Computing Systems)*, Austin, Texas, 227–233.

Hollan, James; Rich, Elaine; Hill, William; Wroblewski, David; Wilner, Wayne; Wittenburg, Kent; Grudin, Jonathan; and members of the Human Interface Lab (1991). "An introduction to HITS: Human Interface Tool Suite." In *Intelligent User Interfaces*, edited by Joseph W. Sullivan and Sherman W. Tyler, 293–338. ACM Press and Addison-Wesley.

Sullivan, Joseph W., and Tyler, Sherman W., eds. (1991). *Intelligent User Interfaces*. ACM Press and Addison-Wesley.

Wroblewski, David; McCandless, Timothy; and Hill, William (1991) "DETENTE: Practical support for practical action." In *Proceedings, CHI '91 (Conference on Human Factors in Computing Systems)*, New Orleans, Louisiana, 195–202.

Kent Wittenburg is a member of technical staff in the Computer Graphics and Interactive Media Research Group at Bellcore, where he works on theory and applications of multidimensional parsing in interfaces. He is founder and coordinator of the ACL Special Interest Group on Multimedia Language Processing (SIGMEDIA). Wittenburg's address is: Bellcore, 445 South St., Room MRE 2A-347, Morristown, NJ 07962-1910; e-mail: kentw@bellcore.com