In: ECDL'98, Lecture Notes in Computer Science, Vol. 1513, Heidelberg: Springer, 1998, pp. 1-11.

Intelligent Multimedia Communication

Mark Maybury[1], Oliviero Stock[2], and Wolfgang Wahlster[3]

[1]The MITRE Corporation
202 Burlington Road, Bedford, MA 01730, USA
maybury@mitre.org

[2]Istituto per la Ricerca Scientifica e Tecnologica (IRST)
via Sommarive 18,I-38050 Povo - Trento, Italy
stock@irst.itc.it

[3]German Research Center for AI (DFKI)
Stuhlsatzenhausweg 3, D-66119 Saarbruecken, Germany
wahlster@dfki.uni-sb.de

Abstract. Multimedia communication is a part of everyday life and its appearance in computer applications is increasing in frequency and diversity. Intelligent or knowledge based computer supported communication promises a number of benefits including increased interaction efficiency and effectiveness. This article defines the area of intelligent multimedia communication, outlines fundamental research questions, summarizes the associated scientific and technical history, identifies current challenges and concludes by predicting future breakthroughs including multilinguality. We conclude describing several new research issues that systems of systems raise.

1 Definition of Multimedia Communication

We define *communication* as the interaction between human-human, human-system, and human-information. This includes interfaces to people, interfaces to applications, and interfaces to information. Following Maybury and Wahlster [1], we define:

- *Multimedia* - physical means via which information is input, output and/or stored (e.g., interactive devices such as keyboard, mouse, displays; storage devices such as disk or CD-ROM)
- *Multimodal* - human perceptual processes such as vision, audition, taction
- *Multicodal* - representations used to encode atomic, elements, syntax, semantics, pragmatics and related data structures (e.g., lexicons, grammars) associated with media and modalities.

The majority of computational efforts have focused on multimedia human computer interfaces. There exists a large literature and associated techniques to develop learnable, usable, transparent interfaces in general (e.g., Baecker et al. [2]). In particular, we focus here on intelligent and multimedia user interfaces (e.g.,Maybury [3]) which from the user perspective assist in tasks, are context sensitive, adapt appropriately (when, where, how) and may:

- *Analyze* sychronous and asychronous multimedia/ modal input (e.g., spoken and written text, gesture, drawings) which might be imprecise, ambiguous, and/or partial
- *Generate* (design, realize) coordinated, cohesive, and coherent multimedia/modal presentations
- *Manage* the interaction (e.g., training, error recovery, task completion, tailoring interaction) by representing, reasoning, and exploiting *models* of the domain, task, user, media/mode, discourse, and environment.

From the developers perspective there is also interest in decreasing the time, expense, and level of expertise necessary to construct successful systems. Finally, when interacting with information spaces, there is the area of media content analysis (Maybury [4]) which includes retrieval of text, audio, imagery and/or combinations thereof.

2 Fundamental Questions

The fundamental questions mirror the above definitions:

- *Analysis*: How do we build systems to deal with synchronous and asychronous, imprecise, ambiguous, and/or partial multimedia/modal input?
- *Generation*: How do we design, realize, and tailor coordinated, cohesive, and coherent multimedia/modal presentations?
- *Management*: How do we ensure efficient, effective and natural interaction (e.g., training, error recovery, task completion, tailoring interaction styles)? How do we represent, reason, and exploit models of the domain, task, user, media/mode, and context (discourse, environment)?
- *Methods*: What kinds of representations and reasoning is required to enable the above. What kinds of multimedia corpora are required? What kinds of evaluation measures, metrics and methods will move this area forward?

3 Timeline

There has been interest in computer supported multimedia communication for the past three decades. We briefly characterize the major problems addressed, developments, and influence to related areas. We characterize some landmark developments using the above distinctions of analysis of input, generation of output, and interaction management.

Late 1950s
- Input/Output: Natural language interfaces (NLI) discussed at Dartmouth AI Conference. First integrated graphics/pointing system developed & deployed (SAGE) [5] [6].

1960s
- Input/Output: Laboratory investigations of VR, initial interest in NLI
- General: First Conference on Computational Linguistics [7]

1970s
- Input: Applications of NLI
- Output: Template generation systems. First speech to text systems.
- Management: focus and dialogue coherence models
- General: emerging commercial systems

1980s
- Input: Commercialization of NLI; First integrated speech and gesture (e.g., Bolt's "Put that there" [8]).

- Output: Creation of techniques for domain independent, rhetorically structured text (e.g., rhetorical schemas, communicative plans). Improved sentence/clause planning/realization. First multilingual generation systems. Automated graphics design.
- Management: User and Discourse modeling. Model-based interfaces.
- General: International workshops on user modeling, text generation, multimodal interaction (e.g., VENACO); government programs (e.g., DARPA intelligent user interface program), industrial visions of intelligent multimodal, multilingual interaction (e.g., Apple's "Phil", AT&T).

1990s

- Input: Increasing spoken language applications. More sophisticated input analysis prototypes (e.g., partial, synchronous, and ambiguous input).
- Output: Coordinated, multimodal generation prototypes. Standard reference model for presentation systems.
- Management: User adapted systems. Agents appear in commercial software (e.g., Microsoft™ Office Assistant).
- General: DARPA (e.g., Intelligent Collaboration and Visualization program (http://snad.ncsl.nist.gov/~icv-ewg/, "Communicator" spoken language architecture) and European Community Intelligent Information Interfaces (www.i3net.org) programs. First ACM international conference on intelligent user interfaces (IUI) [9]. Readings in IUI [1]. Emergence of media content analysis for new applications, e.g., news understanding, video mail and/or VTC indexing and retrieval.

4 Examples of Multimedia Information Access

Significant progress has been made in multimedia interfaces, integrating language, speech, and gesture. For example, Figure 1 shows the CUBRICON system architecture [10]. CUBRICON enables a user to interact using spoken or typed natural language and gesture, displaying results using combinations of language, maps, and graphics. Interaction management is effected via mechanisms such as a user and discourse model which not only influence the generated responses but also manage window layout based on user focus of attention.

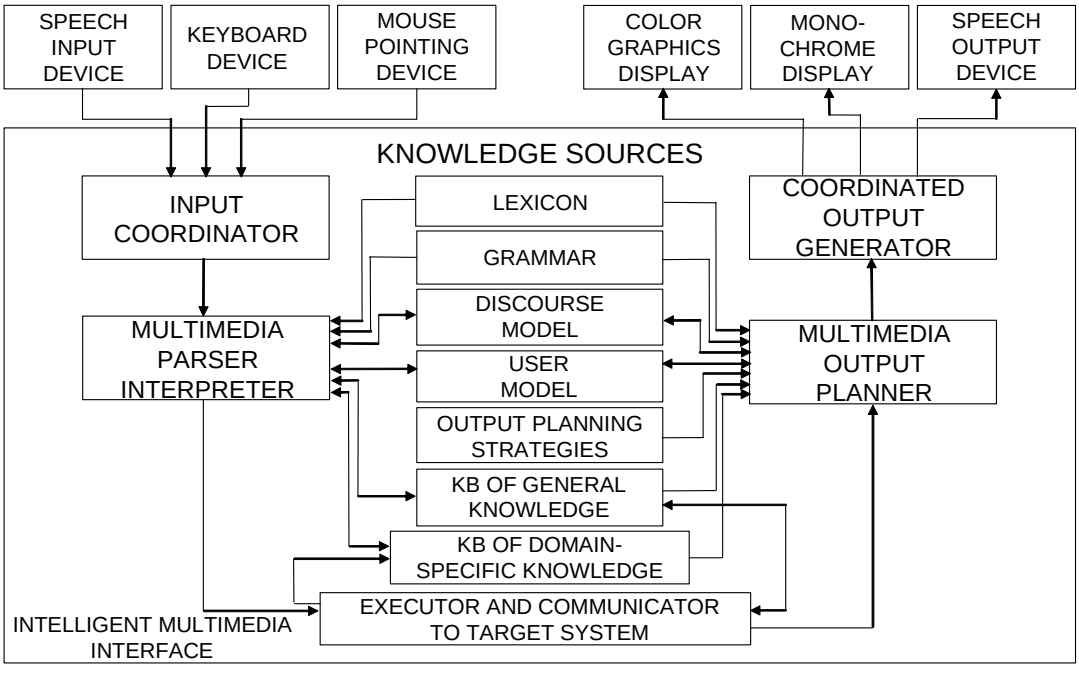

Fig. 1. CUBRICON Multimedia Interface Architecture

In a strongly related example, progress has been made in multimedia information access, integrating language, speech, and image processing, together with more traditional techniques such as hypermedia. For example, Stock et al.'s Alfresco ([11] [12]) is a system for accessing cultural heritage information that integrates in a coherent exploration dialogue a) language based acts, with implicit and explicit reference to what has been said and shown, and b) hypermedia navigation. The generation system, part of the output presentation system, is influenced by the user's interest model that develops in the course of the multimodal interaction. Another aspect developed in this system is cross-modal feedback (Zancanaro et al. [13]). The user is provided fast graphical feedback of the discourse references interpretation by the system, exploiting profitably the large bandwidth of communication we have in a multimodal system.

In a related are of media understanding [4], systems are beginning to emerge that process sychronous speech, text, and images. For example, Figure 2 shows the results of a multimedia news analysis system that exploits redundancy across speech language (closed caption text) and video to mitigate the weaknesses of individual channel analyzers (e.g., low level image analysis and errorful speech transcription). After digitizing, segmenting (into stories and commercials), extracting (pulling out named entities [14], and summarizing into key frames and key sentences, MITRE's BNN [15], enables a user is able to browse and search broadcast news and/or visualizations thereof. A range of on-line customizable views of news summaries by time, topic, or named entity enable the user to quickly retrieve segments of relevant content.

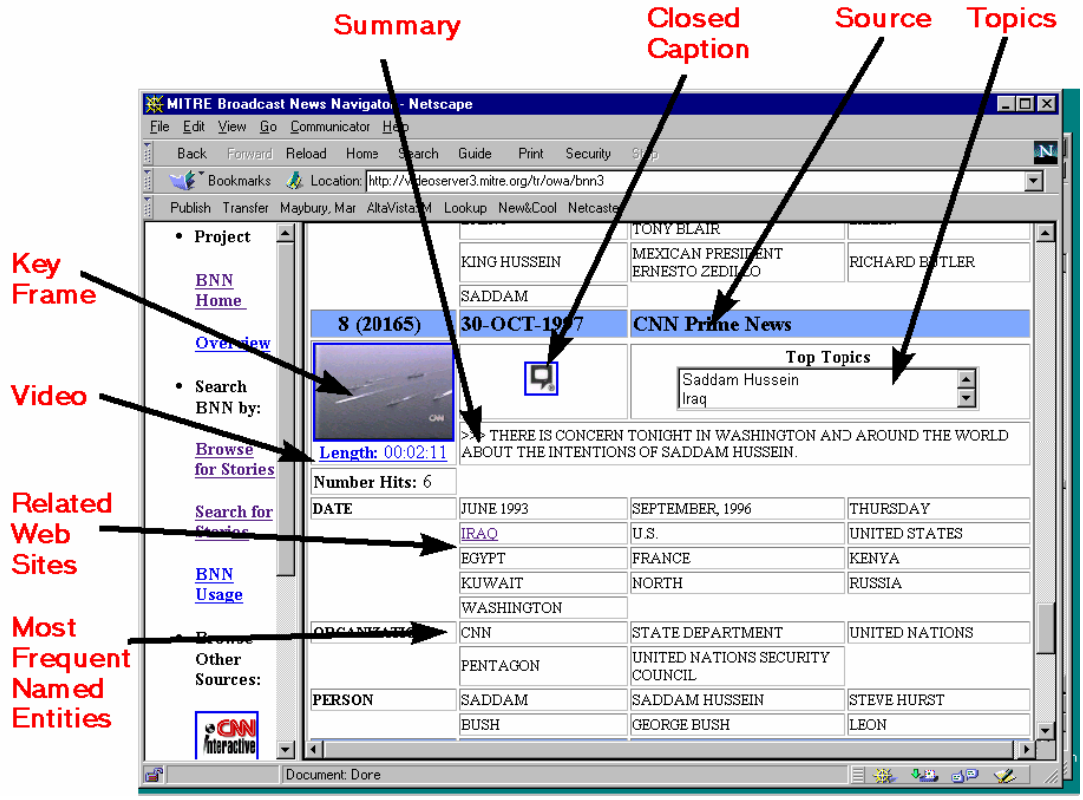

Fig. 2. Detailed Video Story Display

In terms of multilingual information access, one problem is that machine translation systems often provide only gist quality translations that, nonetheless, can be useful to end users in relevancy judgements. Figure 3 illustrates, a web page retrieved from the web by searching for German chemical companies on the web using the German word "chemie" and "gmbh". After retrieving German-language web site, we use a web based machine translation engine (Systran) to obtain a gist-quality translation of their chemical products (Figure 4). Note how the HTML document structure enhances the intelligibility of the resultant translation.

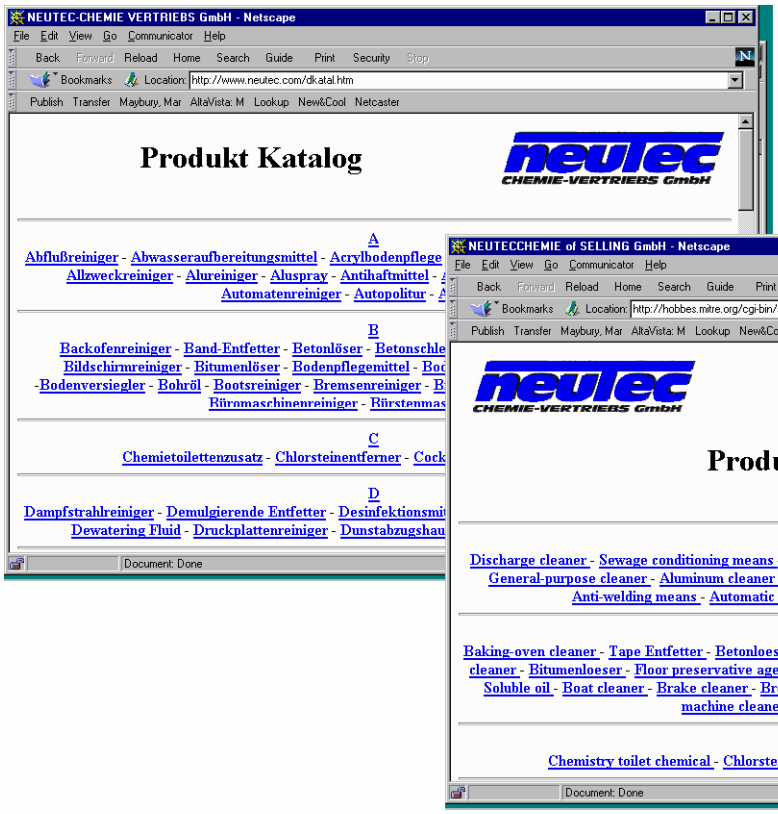

Fig. 3. Original Foreign Language Internet Page

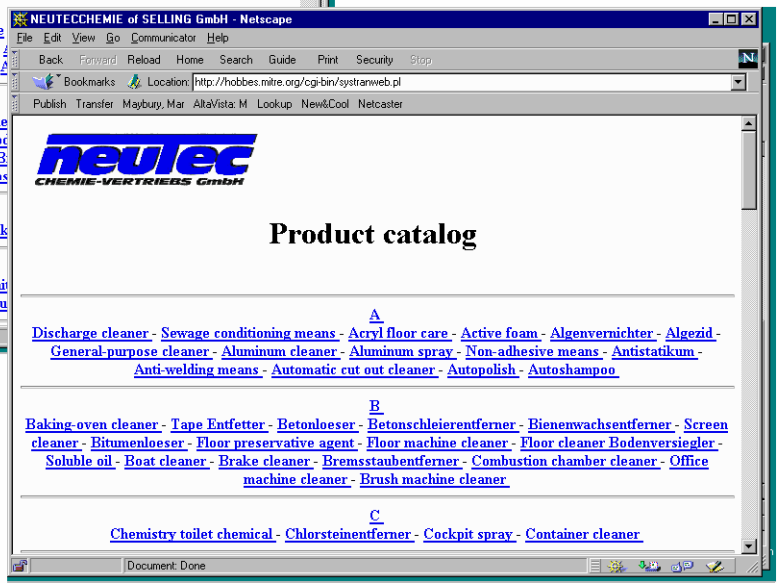

Fig. 4. Translated Language Internet Page

5 Impediments to Progress

Research is not advancing as rapidly as possible in this area because of several impediments. These include:

1. The need for media specific and media independent representations (e.g., a graphical lexicon, syntax, semantics and perhaps even pragmatics and its relation to those in language and gesture), as well as associated tools and techniques that foster reuse and provide a shared foundation for community research.

2. The creation and sharing of resources [16], in particular Multi-* corpora, that is:
   - Multimedia content (e.g., Web, News, VTC)
   - Multimedia interaction (need for instrumentation)
   - Multiparty interaction (e.g., CSCW)

   Found data might include chat sessions and multilingual radio and television corpora and/or multilingual websites.

3. Standards and system modules for plug-and-play evaluation
   - Interfaces: distributed computing infrastructure with reusable elements including speech and language modules, user models (e.g., BGP-MS) & discourse modules, agent communication and coordination (e.g., KQML, open agent architectures)
   - Media analyzers
   - Intellectual property, e.g., ownership and distribution of media and knowledge resources

6 Major Breakthroughs Coming

In the next five years, current research will likely yield several key outcomes. Areas of expected advancement include:

- Integration of language processing and hypermedia
- Integration of multimodal processing mechanisms, e.g., image and language processing
- Increasing transition of advances in agent technology into commercial interface applications
- Transfer of HCI evaluation techniques (e.g., wizard of oz, cognitive walkthrough, task-based evaluation) to multimodal communication research.

Given advances in corpus based techniques for information retrieval and information extraction (e.g., commercial tools for proper name extraction with 90% precision and recall), and the current transfer of these techniques to multilingual information retrieval and extraction, we can expect their application to multilingual understanding. We also believe there is an equivalent opportunity for multimedia generation for other languages.

Transfer of techniques from related areas will be an important strategy. For example, researchers are beginning to take statistical and corpus based techniques formerly applied to single media (e.g., speech, text processing) and apply these to multimedia (e.g., VTC, TV, CSCW).

This work will enable new application areas for relatively unexplored tasks including:

- Multimodal/lingual information access
- Multimodal/lingual presentation generation (summarization)
- Multimodal/lingual collaboration environments

A number of fundamental issues will need to be addressed, such as those outlined below.

7 Role of Multiple Languages

As indicated above, multimodal interaction is the integration of multiple subfields. When extending techniques and methods to multiple languages, we have the benefit of drawing upon previous monolingual techniques that have generality. For example, many generation techniques and components (e.g., content selection, media allocation, media design) built previously for monolingual generation, can mostly be reused across languages. Analogously, interaction management components (e.g., user and discourse models) can be reused. There are, of course, many language specific phenomena that need to be addressed. For example, in generation of multilingual and multimedia presentations, lexical length affects the layout of material both in space and in time. For instance, in laying out a multilingual electronic yellow pages, space may be strictly limited given a standard format and so variability in linguistic realization across languages may pose challenges. In a multimedia context, we might need to not only generate language specific expressions, but also culturally appropriate media.

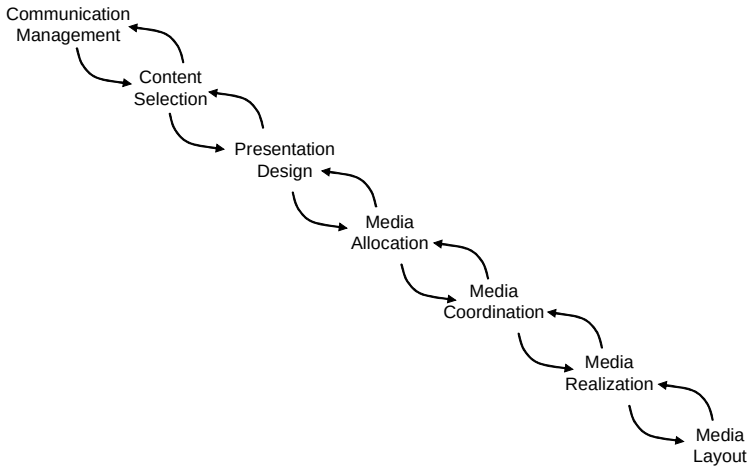

Figure 5. Generation Tasks

In making further progress in this area, there are some unique resources we might take advantage to help develop systems for multimedia information access (e.g., dubbed movies, multilingual broadcast news) that might help accelerate the development of, for example, a multilingual video corpora.

8 Systems Research

Multimedia communication systems, which incorporate multiple subsystems for analysis, generation and interaction management, raise new research questions beyond the well known challenges which occur in component technologies (e.g., learnability, portability, scalability, performance, speed within a language processing system). These include intersystem error propagation, intersystem control and invocation order, and human system interaction.

8.1 Evaluation and Error Propagation

As systems increasingly integrate multiple interactive subcomponents, there is an opportunity to integrate and/or apply software in parallel or in sequence to a given information source or sink. For example, in an information access application where the user may retrieve, extract, translate, or summarize information, we have the possibility of influencing the utility of the output just by sequencing systems given their inherent performance properties (e.g., accuracy, speed). For example, we might use language processing to enhance post-retrieval analysis (extract common terms across documents, re-rank documents provide translated summaries) to hone in on relevant documents. These documents might then cue the user to effective keywords to search for foreign language sources, whose relevancy is assessed using a fast but low quality web-based translation engine. Translating content initially would have been costly, slow, and ineffective. An analogous situation arises when searching multimedia repositories. Old and new evaluation measures, metrics, and methods will be required in this multifaceted environment.

8.2 Multilingual and Multimodal Sources

New research opportunities are raised by processing multilingual and multimodal sources, including the challenge of summarizing across these. For example, what is the optimal presentation of content and in which media or mix of media? [17]. Or consider that in broadcast news spoken language transcription, the best word error rates are currently around 10% for broadcast news anchor speech. What is the cascaded effect of subsequently extracting entities, summarizing, or translating the text?

This also extends to the nature of the interface with the user. For example, applying a low quality speech-to-text transcriber followed by a high quality summarizer may actually result in poorer task performance than providing the user with rapid auditory preview and skimming of the multimedia source.

8.3 User Involvement in Process?

How should users interact with these language-enabled machines? Users of Alta-Vista are now shown foreign web sites matching their queries, with offers to translate them. When invoking the translator, however, the user must pick the source and target language, but what if the character sets and language are unrecognizable by the user? What kind of assistance should the user provide the machine and vice versa? Should this extend to providing feedback to enable machine learning? Would this scale up to a broad set of web users? An in terms of multimedia interaction, who do we develop models of interaction that adequately address issues such as uni- and multi-modal (co)reference, ambiguity, and incompleteness?

8.4 Resource Inconsistencies

Finally, with the emergence of multiple language tools, users will be faced with systems which use different language resources and models. This can readily result in incoherence across language applications, the most obvious case being when a language analysis module is able to interpret a user query containing a given word but a separate language generation module chooses a different word because the original is not in its vocabulary, resulting in potential undesired implicatures by the user. For example, if a user queries a multilingual database in English for documents on "chemical manufacturers" and this is translated into a query for "chemical companies", many documents on marketing and distribution companies would also be included. If these were then translated and summarized, a user might erroneously infer most chemical enterprises were not manufacturers. The situation can worsen above the lexical level when dealing with user and discourse models which are inconsistent across applications.

9 Conclusion

We have outlined the history, developments and future of systems and research in multimedia communication. If successfully developed and employed, these systems promise:

- More *efficient* interaction -- enabling more rapid task completion with less work.
- More *effective* interaction -- doing the right thing at the right time, tailoring the content and form of the interaction to the context of the user, task, dialogue
- More *natural* interaction -- supporting spoken, written, and gestural interaction, ideally as if interacting with a human interlocutor, but taking also into account the potentially extended bandwidth of communication

Because of the multidimensional nature of multimedia communication, interdisciplinary teams will be necessary and new areas of science may need to be invented (e.g., moving beyond psycholinguistic research to "psychomedia" research). New, careful theoretical and empirical investigations as well as standards to ensure cross system synergy will be required to ensure the resultant systems will enhance and not detract from the cognitive ability of end users.

References

1. Maybury, M. T. and Wahlster, W. editors. *Readings in Intelligent User Interfaces.* Morgan Kaufmann Press. ISBN: 1-55860-444-8. (1998)
2. Baecker, R.; Grudin, J.; Buxton, W.; and Greenberg, S. second edition. *Readings in Human-Computer Interaction: Toward the Year 2000.* San Francisco: Morgan Kaufmann. (1995)
3. Maybury, M. T. editor. *Intelligent Multimedia Interfaces.* AAAI/MIT Press. ISBN 0-262-63150-4. (http:/www.aaai.org:80/Press/Books/Maybury1/maybury.html) (1993)
4. Maybury, M. T. editor. *Intelligent Multimedia Information Retrieval.* AAAI/MIT Press. (http:/Avw.aaai.org:80/Press/Books/Maybury2) (1997)
5. Everett, R. et al. 1957. "SAGE: A Data-Processing System for Air Defense," In Proceedings of the Eastern Joint Computer Conference, Washington, D.C., December, 1957.
6. Everett, R. et al. 1983. "SAGE: A Data Processing System for Air Defense, *"Annals of the History of Computing*, 5(4) October 1983.
7. http://www.aclweb.org/
8. Bolt, R. A. "Put-That-There": Voice and Gesture at the Graphics Interface. *ACM Computer Graphics* 14(3): 262-270. Quarterly Report of SIGGRAPH-ACM SIGGRAPH'80 Conference Proceedings, July 14-18. Seattle, Washington. (1980)
9. http://sigart.acm.org:80/iui99/
10. Neal, J. G.; Thielman, C. Y.; Dobes, Z.; Haller, S. M. and Shapiro, S. C. Natural Language with Integrated Deictic and Graphic Gestures. In *Proceedings of the 1989 DARPA Workshop on Speech and Natural Language*, 410-423, Harwich Port: Morgan Kaufmann (1989)
11. Stock. O. and the NLP Group. AlFresco: Enjoying the Combination of NLP and Hypermedia for Information Exploration. In M. Maybury (ed.) *Intelligent Multimedia Interfaces*, AAAI Press, Menlo Park, CA. (1993)
12. Stock, O., Strapparava, C. and Zancanaro, M. Explorations in an Environment for Natural Language Multimodal Information Access. In M. Maybury (ed.) *Intelligent Mutimodal Information Retrieval.* AAAI Press, Menlo Park, Ca./MIT Press, Cambridge, MA. (1997)
13. Zancanaro, M., Stock, O. and Strapparava, C. Multimodal Interaction for Information Access: Exploiting Cohesion. In *Computational Intelligence*, 13 (4). (1997)
14. Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P., and Vilain, M. "Description of the Alembic System Used for MUC-6," Proceedings of the Sixth Message Understanding Conference. Advanced Research Projects Agency Information Technology Office, Columbia, MD, 6-8 November (1995)
15. Merlino, A., Morey, D., and Maybury, M. "Broadcast News Navigation using Story Segments", ACM International Multimedia Conference, Seattle, WA, November 8-14, 381-391.(1997)
16. First International Conference on Language Resources and Evaluation (LREC). Workshop on Multilingual Information Management, Granada, Spain. May 31-June 1 (1998) pp 68-71.
17. Merlino, A. and Maybury, M. An Empirical Study of the Optimal Presentation of Multimedia Summaries of Broadcast News. In Mani, I. and Maybury, M. (eds.) *Automated Text Summarization.* (to appear)