PLANNING MULTIMODAL DISCOURSE

Wolfgang Wahlster
German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3
D-6600 Saarbrücken 11, Germany
Internet: wahlster@dfki.uni-sb.de

Abstract

In this talk, we will show how techniques for planning text and discourse can be generalized to plan the structure and content of multimodal communications, that integrate natural language, pointing, graphics, and animations. The central claim of this talk is that the generation of multimodal discourse can be considered as an incremental planning process that aims to achieve a given communicative goal.

One of the surprises from our research is that it is actually possible to extend and adapt many of the fundamental concepts developed to date in computatational linguistics in such a way that they become useful for multimodal discourse as well. This means that an interesting methodological transfer from the area of natural language processing to a much broader computational model of multimodal communication is possible. In particular, semantic and pragmatic concepts like speech acts, coherence, focus, communicative act, discourse model, reference, implicature, anaphora, rhetorical relations and scope ambiguity take an extended meaning in the context of multimodal discourse.

It is an important goal of this research not simply to merge the verbalization and visualization results of mode-specific generators, but to carefully coordinate them in such a way that they generate a multiplicative improvement in communication capabilities. Allowing all of the modalities to refer to and depend upon each other is a key to the richness of multimodal communication.

A basic principle underlying our model is that the various constituents of a multimodal communication should be generated from a common representation of what is to be conveyed. This raises the question of how to decompose a given communicative goal into subgoals to be realized by the mode-specific generators, so that they complement each other. To address this problem, we explore computational models of the cognitive decision process, coping with questions such as what should go into text, what should go into graphics, and which kinds of links between the verbal and non-verbal fragments are necessary. In addition, we deal with layout as a rhetorical force, influencing the intentional and attentional state of the discourse participants.

We have been engaged in work in the area of multimodal communication for several years now, starting with the HAM-ANS (Wahlster et al. 1983) and VITRA systems (Wahlster 1989), which automatically create natural language descriptions of pictures and image sequences shown on the screen. These projects resulted in a better understanding of how perception interacts with language production. Since then, we have been investigating ways of integrating tactile pointing and graphics with natural language understanding and generation in the XTRA (Wahlster 1991) and WIP projects (Wahlster et al. 1991).

The task of the knowledge-based presentation system WIP is the context-sensitive generation of a variety of multimodal communications from an input including a presentation goal (Wahlster et al. 1993a). The presentation goal is a formal representation of the communicative intent specified by a back-end application system. WIP is currently able to generate simple multimodal explanations in German and English on using an espresso machine, assembling a lawn-mower, or installing a modem, demonstrating our claim of language and application independence. WIP is a highly adaptive multimodal presentation system, since all of its output is generated on the fly and customized for the intended discourse situation. The quest for adaptation is based on the fact that it is impossible to anticipate the needs and requirements of each potential dialog partner in an infinite number of discourse situations. Thus all presentation decisions are postponed until runtime. In contrast to hypermedia-based approaches, WIP does not use any preplanned texts or graphics. That is, each presentation is designed from scratch by reasoning

from first principles using commonsense presentation knowledge.

We approximate the fact that multimodal communication is always situated by introducing seven discourse parameters in our model. The current system includes a choice between user stereotypes (e.g. novice, expert), target languages (German vs. English), layout formats (e.g. paper hardcopy, slide, screen display), output modes (incremental output vs. complete output only), preferred mode (e.g. text, graphics, or no preference), and binary switches for space restrictions and speech output. This set of parameters is used to specify design constraints that must be satisfied by the final presentation. The combinatorics of WIP's contextual parameters can generate 576 alternate multimodal presentations of the same content.

At the heart of the multimodal presentation system WIP is a parallel top-down planner (André and Rist 1993) and a constraint-based layout manager. While the root of the hierarchical plan structure for a particular multimodal communication corresponds to a complex communicative act such as describing a process, the leaves are elementary acts that verbalize and visualize information specified in the input from the back-end application system.

In multimodal generation systems, three different processes are distinguished: a content planning process, a mode selection process and a content realization process. A sequential architecture in which data only flow from the "what to present" to the "how to present" part has proven inappropriate because the components responsible for selecting the contents would have to anticipate all decisions of the realization components. This problem is compounded if content realization is done by separate components (e.g. for language, pointing, graphics and animations) of which the content planner has only limited knowledge.

It seems even inappropriate to sequentialize content planning and mode selection. Selecting a mode of presentation depends to a large extent on the nature of the information to be conveyed.

On the other hand, content planning is strongly influenced by previously selected mode combinations. E.g., to graphically refer to a physical object (Rist and André 1992), we need visual information that may be irrelevant to textual references. In the WIP system, we interleave content and mode selection. In contrast to this, presentation planning and content realization are performed by separate components to enable parallel processing (Wahlster et al. 1993b).

In a follow-up project to WIP called PPP (Personalized Plan-Based Presenter), we are currently addressing the additional problem of planning presentation acts such as pointing and coordinated speech output during the display of the multimodal material synthesized by WIP.

The insights and experience we gained from the design and implementation of the multimodal systems HAM-ANS, VITRA, XTRA and WIP provide a good starting point for a deeper understanding of the interdependencies of language, graphics, pointing, and animations in coordinated multimodal discourse.

REFERENCES

André, Elisabeth; and Rist, Thomas. 1993. The Design of Illustrated Documents as a Planning Task. Maybury, Mark (ed.). *Intelligent Multimedia Interfaces*, AAAI Press (to appear).

Rist, Thomas; and André, Elisabeth. 1992. From Presentation Tasks to Pictures: Towards an Approach to Automatic Graphics Design. Proceedings European Conference on AI (ECAI-92), Vienna, Austria (1992) 764-768.

Wahlster, Wolfgang. 1989. One Word Says more than a Thousand Pictures. On the Automatic Verbalization of the Results of Image Sequence Analysis Systems. *Computers and Artificial Intelligence*, 8, 5: 479-492

Wahlster, Wolfgang. 1991. User and Discourse Models for Multimodal Communication. in: Sullivan, J.W.; and Tyler, S.W.(eds.). *Intelligent User Interfaces*, Reading: Addison-Wesley (1991): 45-67.

Wahlster, Wolfgang; Marburger, Heinz; Jameson, Anthony; Busemann, Stephan. 1983. Overanswering Yes-No Questions: Extended Responses in a NL Interface to a Vision System. *Proceedings of IJCAI-83*, Karlsruhe: 643-646.

Wahlster, Wolfgang; André, Elisabeth; Graf, Winfried; and Rist, Thomas. 1991. Designing Illustrated Texts: How Language Production is Influenced by Graphics Generation. *Proceedings European ACL Conference*, Berlin, Germany: 8-14.

Wahlster, Wolfgang; André, Elisabeth; Bandyopadhyay, Som; Graf, Winfried; and Rist, Thomas. 1993a. WIP: The Coordinated Generation of Multimodal Presentations from a Common Representation, in: Ortony, A.; Slack, J.; and Stock, O.(eds.). *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues*, Springer: Heidelberg: 121-144.

Wahlster, Wolfgang; André, Elisabeth; Finkler, Wolfgang; Profitlich, Hans-Jürgen; and Rist, Thomas. 1993b. Plan-Based Integration of Natural Language and Graphics Generation. *Artificial Intelligence Journal* 26(3), (to appear).