Pointing, Language and the Visual World:

Towards Multimodal Input and Output for Natural Language Dialog Systems

Introduction to a panel chaired by

Wolfgang Wahlstcr
Computer Science Department
University of Saarbriicken
6600 Saarbriicken 11, FRG

Panelists: Jtirgen Allgayer, Erhard W. Hinrichs, Jaap Ph. Hoepelman, Willem Levelt, Norman Sondheimer

In face-to-face conversation humans frequently use *deictic gestures* (e.g. the index finger points at something) in parallel to verbal descriptions for referent identification. Such a *multimodal* mode of communication can improve human interaction with machines, as it simplifies and speeds up reference to objects in a visual world.

The basic technical prerequisites for the integration of pointing and natural language (NL) are fulfilled (high-resolution bit-mapped displays and window systems for the presentation of visual information, various pointing devices such as light-pen, mouse, and touch-sensitive screens for deictic input). But the remaining AI problem is that explicit meanings must be given to natural pointing behavior in terms of a formal semantics of the visual world.

Unlike the usual semantics of mouse clicks in direct manipulation environments, in human conversation the region at which the user points (the *demonstratum)* is not necessarily identical with the region which he intends to refer to (the *referent).* In conventional systems there exists a simple one-to-one mapping of a demonstratum onto a referent, and the reference resolution process does not depend on the situational context. Moreover, the user is not able to control the granularity of a pointing gesture, since the size of the predefined mouse-sensitive region specifies the granularity.

Compared to that, natural pointing behavior is much more flexible, but also possibly ambiguous or vague. Without a careful analysis of the discourse context of a gesture there would be a high risk of reference failure, as a deictic operation does not cause visual feedback from the referent (e.g. inverse video or blinking as in direct manipulation systems).

Although the * common visual world' of the user and the system could be any graphics or image, current projects combining pointing and natural language focus on forms or geographic maps.

For example, the TACTILUS subcomponent of our XTRA system handles a variety of tactile gestures, including different granularities, inexact pointing gestures, and pars-pro-toto deixis. In the latter case, the user points at an embedded region when actually intending to refer to a superordinated region. XTRA provides NL access to an expert system, which assists the user in filling out a tax form. During the dialog, the relevant page of the tax form is displayed on one window of the screen, so that the user can refer to regions of the form by tactile gestures. The syntax and semantics of the tax form is represented as a directed acyclic graph (including relations such as 'geometrically embedded' or 'conceptual part of), which contains links to concepts in a KL-ONE knowledge base.

The deixis analyzer of XTRA is realized as a constraint propagation process over these networks. In addition, TACTILUS uses various other knowledge sources of XTRA (e.g. the semantics of the accompanying verbal description, case frame information, the dialog memory) for the interpretation of the pointing gesture.

While the simultaneous exploitation of both verbal and non-verbal channels provides maximal efficiency, most of the current prototypes don't use truly parallel input techniques, since they combine *typed* NL and pointing. In these systems the user's hands move frequently back-and-forth from the keyboard to the pointing device. Note however, that multimodal input makes even NL interfaces without speech input more acceptable (less keystrokes) and that the research on typed NL forms the basis for the ultimate speech understanding system.

Another restriction of current prototypes is that the presented visual material is fixed and finite, so that the system builder can encode its semantics into the knowledge base. While some of the recent NL interfaces respond to queries by generating graphics, they are not able to analyze and answer follow-up questions about the form and content of this graphics, since they do not have an appropriate representation of its syntax and semantics. Here one of the challenging problems is the automatic formalization of synthetic visual information as a basis for the interpretation of gestural input.

Some of the open questions addressed by the panel are:

- How can non-verbally communicated information be included in a formal semantic representation of discourse?
- What is an adequate architecture of parsers and generators for multimodal communication?
- What effects have gestures on the attentional state and intentional structure of a dialog?
- How could a generator decide whether to use a pointing gesture, a verbal description or a combination of both for referent identification (knowledge-based media choice)?
- What are the temporal interdependences of verbal and non-verbal output in deictic expressions (synchronization of speech and gesture)?
- How can we cope with complex pointing actions, e.g. a continuous movement of the index finger (drawing a circle around a group of objects, underlining something, specifying a direction or a path) or a quick repetition of discrete pointing acts (emphatic pointing, multiple reference)?