Referring to Displays in Multimodal Interfaces

Daqing He **Graeme Ritchie**
Dept. of Artificial Intelligence,
University of Edinburgh,
80 South Bridge,
Edinburgh EH1 1HN, Scotland
daqingd,graeme@dai.ed.ac.uk

John Lee
Human Communication Research Centre,
University of Edinburgh
2 Buccleuch Place,
Edinburgh EH8 9LW, Scotland
john@cogsci.ed.ac.uk

Abstract

A system which displays information graphically, and also allows natural language queries, should allow these queries to interrogate the displayed (visual) information. Ideally this would use some uniform method for processing queries both about the display and about the world model. Such a system would have to cope with ambiguities introduced by these two sources of information. These ambiguities, and a preliminary proposal for a system to deal with it, are the main topics of this paper.

1 Introduction

Projects which have attempted to integrate natural language (NL) with graphical displays (Bès and Guillotin, 1992; Neal and Shapiro, 1991; Pineda, 1989) have mainly focussed on one of two problems:

1. How can output text be coordinated with graphical information displayed on the screen?
2. How can pointing gestures be coordinated with NL input?

We are interested in a slightly different issue, namely:

> How can NL terms be used, in a relatively uniform way, to refer to visual objects on the screen as well as the objects (for example, database items) which they may denote?

The situation we have in mind is where the computer system has some stored knowledge base, database or model, and is able to graphically display selected items from that store. The user wishes to interact with the system, and may wish to ask questions which either allude to visual features of the display (e.g. *Is the blue zone inside the city boundary?*) or are directly about the meaning of the display (e.g. *What does the blue marking represent?*). Such queries require that the system have access to some representation of what is represented on the screen, and that this representation be amenable to NL or multimodal (MM) querying.[1]

2 World Model and Display Model

It is common in systems which present visual information on the screen (e.g. GISs) for there to be a *display model*. This is an explicit representation of what items are currently on the screen and what their characteristics are. This is distinct from the *world model* which represents the facts about the world that the system has, which may not be displayed on the screen. In such systems, the main role of the display model is to maintain the visual display in an orderly fashion, and to connect screen objects to world (or database) objects. It must be updated systematically as items appear, disappear or move on the screen. Very often, the display model is quite a low-level structure, as it performs basic housekeeping for the display.

Our proposal is that, for NL querying of the visual display to be possible, the display model must contain suitable high level information in a form which is accessible to an NL front-end; preferably, this form would be similar to, or related to, the representation the NL front-end uses to access the world model.

3 Illustrative Examples

A non-spatial domain

It might seem that queries about the visual display would make sense only in a domain where *spatial* in-

[1] We shall discuss natural language, but with the assumption that working systems in a few years' time would operate with speech input.

formation is directly relevant, such as a street map or room plan. However, if an iconic display is being used to represent some non-spatial set of objects, it might still be desirable to use visual attributes to refer to these abstract icons. To make these remarks slightly more concrete, let us consider a (fictitious) example system. This system does not handle spatial information, but it uses iconic representations on the screen to convey database facts to the user.

The application is a car-sales catalogue, in which a number of (presumably used/second-hand/pre-owned) cars are available for the user to browse through. Icons on the screen represent individual cars, and various characteristics of the icons convey attributes of the corresponding cars (Figure 1). The

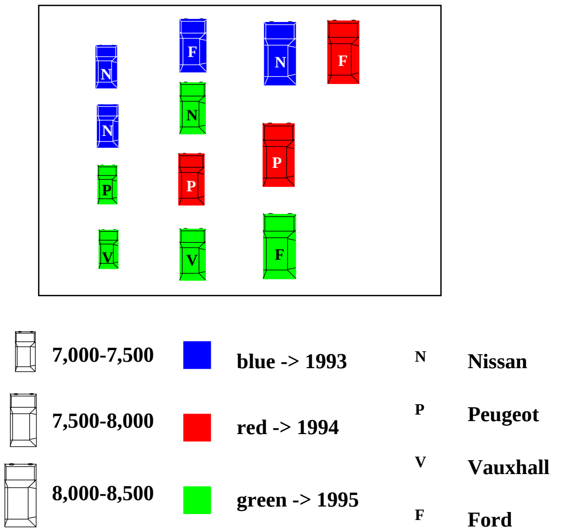

Figure 1: *A car selection system*

size of an icon conveys the price band, the colour conveys the year of production, and the letter on each icon indicates the initial of the manufacturer. The user can point to icons, move icons round, or ask questions about them, such as *What is the insurance group of the car in the top right hand corner?*, *Is the green car a hatchback?*. Notice that spatial phrases (*in the top right hand corner*) can be used, even though this is a non-spatial domain. Also, the colour adjective *green* would (given the coding in Figure 1) probably refer to the colour of the icon on the screen rather than the colour of the actual car, but a similar scenario can be imagined where a colour term would be used to denote the colour of the actual world object. In some cases, both might be possible, leading to ambiguity (Wilson and Conway, 1991).

What is clear from this is that the mapping from the world model (database) to the display model is centrally important. In particular, if we wish to be able to handle questions which are explicitly about the visual representation, such as *What does green represent?*, the mapping itself must be accessible to some form of symbol querying by the NL/MM interface.

A spatial domain

Let us now consider a (fictitious) spatial domain. In this domain, a 2D graphic display is being used to help the user plan the layout of a room. The display represents the overall plan, and icons are stylised images of furnishings and fittings. In such a situation, the user might pose queries such as *What kind of chair is to the right of the table?*, *Would a cupboard fit above the table?*. Here, spatial relations are again used, but there is potential ambiguity as to whether they refer to relations in the world being modelled, or on the screen. An object might be "above" the table in the image, but "to the left" of it in reality.

4 Levels of ambiguity

As argued above, certain forms of reference (e.g. colour, spatial relations) can be ambiguous between visual characteristics of the display and actual characteristics of the world being modelled. For referring expressions, there are two levels to this ambiguity:

Described referent. When the query interpreter is processing a referring expression, it has to determine in which model – the display model or the world model – the features of the object (e.g. colour, size) are being described, and hence used to indicate the referent. During this process, the objects in the world model and those in the display model should be counted as *different* even in cases where a representation relation exists between them.

There may, as noted above, be ambiguity here, between the two models.

Intended referent. Even if a unique object is determined (a display object such as an icon or a world object such as a database item), it is conceivable that this object is being used as a surrogate to refer to the corresponding object under the mapping relation. This can be illustrated using the "car" domain introduced earlier. In a query such as *What is the price of the blue one?*, the colour *blue* may be (unambiguously) a display feature, indicating a blue icon, but the intended referent (for use in the *price* predicate) is the corresponding *world* object, not the icon (cars have prices, icons do not). Conversely, in a command *Move the 1.5 litre car to the top of the screen*, the noun phrase uses domain attributes to indicate a domain object, but the action of *move* is to operate on the corresponding *display* object. The third, and simplest, possibility is that there is no intervening use of the mapping relation – the described referent is itself the intended referent.

This level of indirection can lead to ambiguity when the noun phrase is viewed in isolation, since the choice of intended referent often needs information from the rest of the sentence, or from the context, to disambiguate it.

The consequence of this added level of ambiguity is that the normal way of considering the "sense" and "reference" of an NL phrase has to be reconsidered. Instead of the usual two-level approach in which a symbolic description (the sense) is evaluated, matched, or otherwise processed to produce a particular set of objects (the reference), we need a three-level approach allowing for sense, described referent, and intended referent. All of these have to be managed systematically, so that the correct relationships are maintained, and utilised, between the various objects.

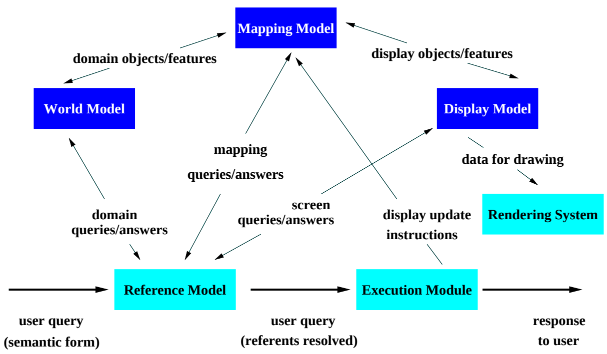

Figure 2: *Proposed architecture for allowing reference to display, world and mapping*

5 Our aim

The aim of our project is to devise a uniform, general and flexible architecture and representation mechanism by which a NL/MM query system can process queries about objects displayed on the screen, about objects in the database or world model, and about the relationship between these two. By "general", we mean that the mechanisms should not be hard-wired or domain-specific. We intend to produce a method whereby a given database and a formally specified visual representation scheme for the database entities can be used to interface directly with a domain independent NL front-end system, thus building a working multi-media system.

6 Preliminary Proposals

In order to facilitate inter-module communication, and to allow for possible symbolic reasoning, a high-level symbolic representation is needed (see figure 2).

Pointing facilities are included in the system, combining with the graphic display and NL interpreter to form a multimodal system. Pointing should assist in resolving ambiguity (of described referent, not intended referent), but the main component for dealing with these ambiguities is the reference model.

It is clear that the resolving of references in such a system could, in general, depend upon a wide variety of sources of knowledge, including the following:

Local Semantic Properties The noun phrase itself will supply the most immediate constraints on choice of (described) referent, in terms of the head noun and its modifiers such as adjectives and prepositional phrases.

Semantic Relations Processing of referents must also take account of relations which are not shown in the noun phrase but which involve the referent(s) and other display/world objects.

Mutual Beliefs The user and the system should know the referent object and its described features, and at the same time both should acknowledge that the other knows the object and its features as well ((Clark and Marshall, 1981), p57). In a multimodal environment, there are various ways for an object to be acknowledged by both dialogue participants: either it is displayed on the screen, or it is mentioned in previous dialogue, or it is part of common sense knowledge for both speaker and listener. A variety of pragmatic inferences might be possible. For example, in the query[2] *What colour is this* $\nearrow$*?*, it may be possible that either model is in question, but it is unlikely that the display property is intended because it is already clearly visible. Moreover, as we suggested, the display property may represent some other property in the world model, so that if the user says *What is the price band of this* $\nearrow$*?*, it may be inferrable both that the user must mean the car (rather than the icon), and also that it would be appropriate to give a reminder (e.g. *Colour represents*

[2]The $\nearrow$ means a pointing act happens here.

price band) about the depictive mapping, since the user is querying a directly depicted world property.

Coherence The coherence of the proceeding dialogue should not be damaged by an object becoming the referent of the expression (Grosz and Sidner, 1986).

It follows that the disambiguation process should be based on the following information sources: *the world model* and *the display model* for the sources of candidates and the examination of various restrictions, *the dialogue model* for providing coherence information about the dialogue and *the user model* for the modelling of mutual beliefs. In practice, our project is too limited to explore all of these issues, and we intend to leave aside issues of mutual belief (that is, our "user model" will be degenerately simple).

It seems plausible that the consideration of described referents could be restricted, in this more limited project, to the use of "Local Semantic Properties" (in the above list). As argued in another context (Ritchie, 1976; Ritchie, 1983), broader semantic constraints (such as relations to other objects or even existence in the current situation) are largely concerned with the eventual referent, rather than superficial aspects of how it happens to be described. Even the question of whether a phrase is a semantically compatible subject or object of a particular verb is a constraint on the referent, not the symbolic expression describing it. In the revised three-level arrangement suggested earlier, such constraints would be on the *intended* referent rather than the *described* referent. That is, in a sentence like *"What kind of fuel-injection system does the blue one have?"* the constraint that the referent must be a type of object which can have a fuel-injection system is to be imposed upon the intended referent.

This suggests, at least superficially, that the described referent might be calculated relatively simply using just the properties of the noun phrase, without much inference. The more difficult question of determining the intended referent would then involve potentially complicated inference about domain objects, etc. This would allow a two stage referent determination approach: find the described referent, then compute possible intended referents. As a benefit of this approach, a pointing action, which can be seen as a short way to indicate a described referent, could be included in a modular fashion.

During these inferences (particularly the search for intended referents), a variety of sources of information may affect the result. It is therefore necessary to have some mechanisms which allow the interaction of these disparate sources. It is possible that some of the constraint-satisfaction suggestions of (Mellish, 1985) might be useful.

If none of the available sources of information resolve the ambiguities, then the query as a whole is ambiguous, but it seems unlikely that this would happen in practice. The challenge is that the processing method should be equally effective at making use of these sources of disambiguation. Our objective is to allow for as much flexibility as we can in the referential phenomena, but we acknowledge inevitable limitations.

Acknowledgements: The first author (Daqing He) is supported by a Colin & Ethel Gordon Scholarship from the University of Edinburgh.

References

Bès, G. and Guillotin, T. (1992). *A natural language and graphics interface, results and perspectives the ACORD project*. Springer-Verlag, Berlin, Germany.

Clark, H. and Marshall, C. (1981). Definite reference and mutual knowledge. In Joshi, A., Webber, B., and Sag, I., editors, *Elements of discourse understanding*, chapter 1, pages 10–63. Cambridge University Press, Cambridge, UK.

Grosz, B. J. and Sidner, C. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204.

Mellish, C. S. (1985). *Computer interpretation of natural language descriptions*. Ellis Horwood series in artificial intelligence. Ellis Horwood.

Neal, J. and Shapiro, S. (1991). Intelligent Multimedia Interface Technology. In Sullivan, J. and Tyler, S., editors, *Intelligent User Interfaces*, pages 11–44. ACM Press.

Pineda, L. A. (1989). *GRAFLOG: A Theory of Semantics fro Graphics with Applications to Human-Computer Interaction and CAD Systems*. PhD thesis, University of Edinburgh, Edinburgh UK.

Ritchie, G. (1976). Problems in local semantic processing. In *Proceedings of AISB Conference*, pages 234–241, Edinburgh, Scotland.

Ritchie, G. (1983). Semantics in parsing. In King, M., editor, *Parsing Natural Language*, pages 199–217. Academic Press, North Holland.

Wilson, M. and Conway, A. (1991). Enhanced interaction style for user interfaces. *IEEE Computer Graphics and Applications*, 11(2):79–90.