On Representing Salience and Reference in Multimodal Human-Computer Interaction

Andrew Kehler[1], Jean-Claude Martin[2], Adam Cheyer[1], Luc Julia[1], Jerry R. Hobbs[1] and John Bear[1]

[1] SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025 USA

[2] LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France

Abstract

We discuss ongoing work investigating how humans interact with multimodal systems, focusing on how successful reference to objects and events is accomplished. We describe an implemented multimodal travel guide application being employed in a set of Wizard of Oz experiments from which data about user interactions is gathered. We offer a preliminary analysis of the data which suggests that, as is evident in Huls et al.'s (1995) more extensive study, the interpretation of referring expressions can be accounted for by a rather simple set of rules which do not make reference to the type of referring expression used. As this result is perhaps unexpected in light of past linguistic research on reference, we suspect that this is not a general result, but instead a product of the simplicity of the tasks around which these multimodal systems have been developed. Thus, more complex systems capable of evoking richer sets of human language and gestural communication need to be developed before conclusions can be drawn about unified representations for salience and reference in multimodal settings.

Introduction

Multimodal systems are particularly appropriate for applications in which users interact with a terrain model that is rich in topographical and other types of information, containing many levels of detail. Applications in this class span the spectrum from travel guide systems containing static, two-dimensional models of the terrain (e.g., a map-based system), to crisis management applications containing highly complex, dynamic, three-dimensional models (e.g., a forest fire fighting system). We are currently investigating how humans interact with multimodal systems in such settings, focusing on how reference to objects and events is accomplished as a user communicates by gesturing with a pen (by drawing arrows, lines, circles, and so forth), speaking natural language, and handwriting with a pen.

In this report, we begin to address the question of how knowledge and heuristics guiding reference resolution are to be represented. Is it possible to have a unified representation for salience that is applicable across multimodal systems, or do new tasks require new representations? Can constraints imposed by the task be modularized in the theory, or are they inherently strewn within the basic mechanisms? Can linguistic theories of reference, which typically treat gestural and spoken deixis as a peripheral phenomenon, be naturally extended to the multimodal case, in which such deixis is the norm?

A Fully Automated Multimodal Map Application

The basis for our initial study is an implemented prototype multimodal travel guide application (Cheyer & Julia 1995) that was inspired by a multimodal Wizard of Oz simulation (Oviatt 1996). The system provides an interactive interface on which the user may draw, write, or speak. The system makes available information about hotels, restaurants, and tourist sites that have been retrieved by distributed software agents from commercial Internet World Wide Web sites.

The types of user interactions and multimodal issues handled can be illustrated by a brief scenario featuring working examples. Suppose Mary is planning a business trip to Toronto, but would like to schedule some activities for the weekend. She turns on her laptop PC, executes a map application, and selects Toronto.

To determine the most appropriate interpretation for the incoming streams of multimodal input, our approach employs an agent-based framework to coordinate competition and cooperation among distributed information sources, working in parallel to resolve the ambiguities arising at every level of the interpretation process. With respect to interpreting anaphora, such as in the command "Show photo of hotel", separate information sources may contribute to the resolution:

- Context by object type: The natural language component can return a list of hotels talked about.
- Deictic: Pointing, circling, or arrow gestures might indicate the referent, which may occur before, during, or after an accompanying verbal command.
- Visual context: The user interface agent might determine that only one hotel is currently visible.

```
M:  [Speaking] Where is downtown?
    Map scrolls to appropriate area.
M:  [Speaking and drawing region]
    Show me all hotels near here.
    Icons representing hotels appear.
M:  [Writes on a hotel] Info?
    A textual description appears.
M:  [Speaking] I only want hotels with a pool.
    Some hotels disappear.
M:  [Draws a crossout on a hotel near a highway]
    Hotel disappears.
M:  [Speaking and circling]
    Show me a photo of this hotel.
    Photo appears.
M:  [Points to another hotel]
    Photo appears.
M:  [Speaking] Price of the other hotel?
    Price appears for previous hotel.
M:  [Speaking and drawing an arrow] Scroll down.
    Display adjusted.
M:  [Speaking and drawing an arrow toward a hotel]
    What is the distance from here to China Town?
    A line and number representing distance displayed.
```

- Database queries: Information from a database agent can be combined with results from other resolution strategies, such as location information for the hotel asked about.
- Discourse analysis: The discourse history provides information for interpreting phrases such as "No, the other one."

The map application is implemented within a multi-agent framework called the Open Agent Architecture (OAA). [3] The OAA provides a general-purpose infrastructure for constructing systems composed of multiple software agents written in different programming languages and running on different platforms. Similar in spirit to distributed object frameworks such as OMG's CORBA or Microsoft's DCOM, agent interactions are more flexible and adaptable than the tightly bound object method calls provided by these architectures, and are able to exploit parallelism and dynamic execution of complex goals. Instead of preprogrammed single method calls to known object services, an agent can express its requests in terms of a high-level logical description of what it wants done, along with optional constraints specifying how the task should be performed. This specification request is processed by one or more Facilitator agents, which plan, execute and monitor the coordination of the subtasks required to accomplish the end goal (Cohen *et al.* 1994).

[3] Open Agent Architecture and OAA are trademarks of SRI International. Other brand names and product names herein are trademarks and registered trademarks of their respective holders.

Application functionality in the map application is thus separated from modality of user interaction. The system is composed of 10 or more distributed agents that handle database access, speech recognition (Nuance Communications Toolkit or IBM's VoiceType), handwriting (by CIC) and gesture (in-house algorithms) recognition, and natural language interpretation. These agents compete and cooperate to interpret the streams of input media being generated by the user. More detailed information regarding agent interactions for the multimodal map application and the strategies used for modality merging can be found in Cheyer and Julia (1995) and Julia and Cheyer (1997).

Data Collection

Despite the coverage of the system's current anaphora resolution capabilities, we are interested in collecting naturally-occurring data which may include phenomena not handled by our system. We therefore designed a Wizard of Oz (WOZ) experiment around the travel guide application. In WOZ experiments, users believe they are interacting directly with an implemented system, but in actuality a human "wizard" intercepts the user's commands and causes the system to produce the appropriate output. The subject interface and wizard interface are depicted in Figure 1.

Experiment Description Subjects were asked to plan activities during and after a hypothetical business trip to Toronto. They planned places to stay, sights to see, and places to dine using speech, writing, and pen-based gestures. The task consisted of four subtasks. To provide experience using each modality in isolation, during the first two tasks subjects planned half days using speech only and pen only respectively. In the third task, subject planned two half-days using any combination of these modalities they wished. Finally, the subjects completed a direction giving task, begun by picking up a phone placed nearby. On the other end was an experimenter who told the subject that he wants to meet for dinner, providing the name of the hotel at which he is staying and the restaurant at which they are to meet. The subject then interacted with the system to determine directions to give to the experimenter. For all tasks, the subjects were given only superficial instruction on the capabilities of the system. The tasks together took an average of 40 minutes. At the end of a session, the subjects were given surveys to determine whether they understood the task and the modalities available to them, and to probe their thoughts on the quality of the system.

The interactions were recorded using video, audio, and computer storage. The video displays a side-by-

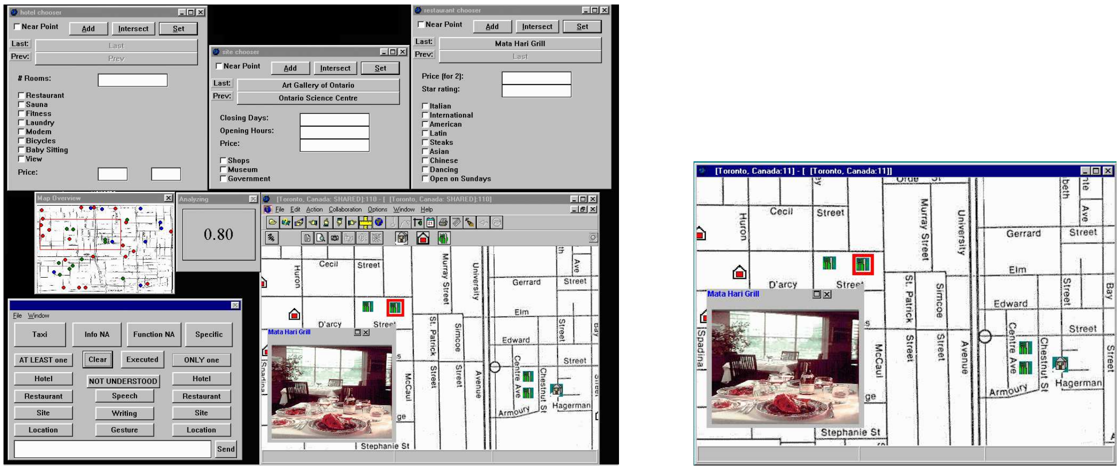

Figure 1: The Wizard Interface (left) and the Subject Interface (right)

side view with the subject on one side and the map interface on the other. The video and audio records are used for transcription, and the computer storage for reenacting scenarios for evaluation.

Coevolution of Multimodal and Wizard-of-Oz Systems In our quest for unconstrained, naturally-occurring data, we sought to place as few assumptions on the user interactions as possible. Unfortunately, WOZ experiments using simulated systems often necessitate such assumptions, so that facilities allowing the wizard to respond quickly and accurately can be encoded. We have improved upon this paradigm by having the wizard use our implemented and highly capable multimodal system to produce the answers to the user.

As described by Cheyer et al. (1998), our multimodal map application already possessed two qualities that allowed it to be used as part of a WOZ experiment. First, the system allows multiple users to share a common workspace in which the input and results of one user may be seen by all members of the session. This enables the Wizard to see the subject's requests and remotely control the display. Second, the user interface can be configured on a per-user basis to include more or fewer graphical user interface (GUI) controls. Thus, the Wizard can use all GUI command options, and also work on the map by using pen and voice. Conversely, the subject is presented with a map-only display. To extend the fully automated map application to be suitable for conducting WOZ simulations, we added only three features: a mode to disable the automatic interpretation of input from the subject, domain-independent logging and playback functions, and an agent-based mechanism for sending WOZ-specific instructions (e.g., Please be more specific.) to the user with text-to-speech and graphics.

The result is a hybrid WOZ experiment: While a naive user is free to write, draw, or speak to a map application without constraints imposed by specific recognition technologies, the hidden Wizard must respond as quickly and accurately as possible by using any available means. In certain situations, a scrollbar or dialog box might provide the fastest response, whereas in others, some combination of pen and voice may be the most efficient way of accomplishing the task. In a single experiment, we simultaneously collect data input from both an unconstrained new user (unknowingly) operating a simulated system – providing answers about how pen and voice are combined in the most natural way possible – and from an expert user (under duress) making full use of our best automated system, which clarifies how well the real system performs and lets us make comparisons between the roles of a standard GUI and a multimodal interface. We expect that this data will prove invaluable from an experimental standpoint, and since all interactions are logged electronically, both sets of data can be applied to evaluating and improving the automated processing.

Performing such experiments and evaluations in a framework in which a WOZ simulation and its corresponding fully functional end-user system are tightly intertwined produces a bootstrap effect: as the automated system is improved to better handle the corpus of subject interactions, the Wizard's task is made easier and more efficient for future WOZ experiments. The methodology promotes an incremental way of designing an application, testing the design through semi-automated user studies, gradually developing the automated processing to implement appropriate behavior for input collected from subjects, and then testing the finished product while simultaneously designing and collecting data on future functionality – all within one unified implementation. The system can also be used without a Wizard, to log data about how real users make use of the finished product.

Data Analysis

At the time of this writing, 17 subjects out of a planned 25 have completed the tasks. We are currently in the process of transcribing and analyzing this data, and so we limit our discussion to a subset of 10 of the sessions. Our conclusions must therefore remain preliminary.

Our analysis of the data covers a broad range of factors concerning modality use. In addition to classical metrics used for analyzing multimodal corpora (monomodal features, temporal relationship between speech and gesture), we are analyzing the commands using a typology based on types of cooperation: specialization, equivalence, redundancy, complementarity, concurrency, and transfer (Martin 1997; Martin, Julia, & Cheyer 1998). Our focus here, however, concerns the use of referring expressions, and we therefore restrict our analysis to this issue.

Models of linguistic reference generally consist of two components. The first is the evolving representation of the discourse state, or "discourse model", which usually includes a representation of the salience of previously introduced entities and events. For instance, entities introduced from an expression occupying subject position are generally considered as being more salient for future reference than those introduced from the direct object or other positions. The second component is a representation of the properties of referring expressions which dictates how they should be interpreted with respect to the discourse model (Prince 1981; Gundel, Hedberg, & Zacharski 1993). For instance, pronouns have been claimed to refer to entities that are highly salient or 'in focus', whereas full definite noun phrases need not refer to salient entities, or even ones that have been mentioned at all. Similarly, the choice among different deictic expressions (i.e., 'this' vs. 'that') is presumably guided by factors relating to the relative places at which their antecedents reside within the discourse model. Within this picture, the representation of discourse state and the interpretation of referring expressions against it are kept distinct; furthermore, they are considered independent of the task underlying the interaction.

An alternative embodied in some multimodal systems, including ours, could be termed the 'decision list' approach. Here, heuristics are encoded as a decision list (i.e., a list of if-then rules applied sequentially) which do not necessarily enforce a strict separation between the representation of multimodally-integrated salience factors and the identities and properties of particular referring expressions. Furthermore, these rules might even query the nature of the task being performed or the type of command being issued, if task analyses would suggest that such differences be accounted for (Oviatt, DeAngeli, & Kuhn 1997).

A unified, modularized theory of reference which is applicable across multimodal applications is presumably preferable to a decision list approach. Huls et al. (1995) in fact take this position and propose such a mechanism. They describe data arising from sessions in which subjects interacted with a system using a keyboard to type natural language expressions and a mouse to simulate pointing gestures. To model discourse state, they utilize Alshawi's (1987) framework, in which *context factors* (CFs) are assigned significance weights and a decay function according to which the weights decrease over time. Significance weights and decay functions are represented together via a list of the form $[w_1,...,w_n,0]$, in which w_1 is an initial significance weight which is then decayed in accordance with the remainder of the list. The *salience value* (SV) of an entity *inst* is calculated as a simple sum of the significance weights $W(CF_i)$:

$$SV(inst) = \sum_{i=1}^{n} W(CF_i^{inst})$$

Four "linguistic CFs" and three "perceptual CFs" were encoded. Linguistic CFs include weights for being in a major constituent position ([3,2,1,0]), the subject position ([2,1,0], in addition to the major constituent weight), a nested position ([1,0]), and expressing a relation ([3,2,1,0]). Perceptual CFs include whether the object is visible ([1,...,1,0]), selected ([2,...,2,0]), and indicated by a simultaneous pointing gesture ([30,1,0]). The weights and decay functions were determined by trial and error.

To interpret a referring expression, the system chooses the most salient entity that meets all type constraints imposed by the command and by the expression itself (e.g., the referent of "the file" in "close the file" must be something that is a file and can be closed). This strategy was used regardless of the type of referring expression. Huls et al. tested their framework on 125 commands containing referring expressions, and compared it against two baselines: (i) taking the most recent compatible reference, and a pencil-and-paper simulation of a focus-based algorithm derived from Grosz and Sidner (1986). They found that all 125 referring expressions were correctly resolved with their approach, 124 were resolved correctly with the Grosz and Sidner simulation, and 119 were resolved correctly with the simple recency-based strategy.

The fact that all of the methods do very well, including a rather naive recency-based strategy, indicates a lack of difficulty in the problem. Particularly noteworthy in light of linguistic theories of reference is that this success was achieved with resolution strategies that were not tied to choice of referring expression. That is, well-known differences between the conditions in which forms such as "it", "this", "that", "here", and "there" are used apparently played no role in interpretation.

We were thus inclined to take a look at the reference behavior shown in our corpus. Table 1 summarizes the distribution of referring expressions within information-seeking commands for our 10 subjects. (Commands to manipulate the environment, such as to scroll the screen or close a window, were not included.) On the vertical axis are the types of referential form used. The symbol ϕ denotes "empty" referring expressions corresponding to phonetically unrealized arguments to commands (e.g., the command "Information", when information is requested for a selected hotel). Full NPs are noun phrases for which interpretation does not require reference to context (e.g., "The Royal Ontario Museum"), whereas definite NPs are reduced noun phrases that do (e.g., "the museum").

On the horizontal axis are categories indicating the information status of referents. We first distinguish between cases in which an object was gestured to (e.g., by pointing or circling) at the time the command was issued, and cases in which there was no such gesture. "Unselected" refers to a (visible) object that is not selected. "Selected Immediate" includes objects that were selected and mentioned in the previous command, whereas "Selected Not Immediate" refers to objects that have remained selected despite intervening commands that have not made reference to it (e.g., due to intervening commands to show the calendar or scroll the screen). There was also one outlying case, in which the user said "Are there any Spanish restaurants here", in which "here" referred to the area represented by the entire map.

These data show a divergence between the distribution of referring expressions and the heuristics one might use to resolve them. On one hand, there are distributional differences in even our admittedly limited amount of data that accord roughly with expectations. For instance, unselected entities, which are presumably not highly salient, were never referred to with pronominal forms without an accompanying gesture. Instead, nonpronominal noun phrases were used (20 full NPs and 2 definite NPs), and in all cases the content of the noun phrase constrained reference to one possible antecedent (e.g., "the museum" when only one museum was visible). Also, the antecedents of empty referring expressions were almost always highly-focused (selected, immediate) objects when no accompanying gesture was used, and "it" always referred to a selected, immediate antecedent. Finally, in accordance with their generally deictic use, "this NPs" (e.g., "this museum") and "this" were usually accompanied by a simultaneous gesture. "Here" was only used when accompanied by such a gesture, whereas "there" was used for all types of selected referents.

Certain other facets of the distribution are more contrary to expectation. For instance, in 36 cases a full NP was used to refer to a selected, immediate object which, as such, was a candidate for a reduced referential expression. In four of these cases, the user also gestured to the antecedent, resulting in an unusually high degree of redundancy. We suspect that such usage may result from a bias some users have regarding the ability of computer systems to interpret natural language.

Despite the distributional differences among the referential forms, a simple algorithm can be articulated which handles all of the data without making reference to the type of referential expression used nor its distributional properties. First, the algorithm narrows the search given any type constraints imposed by the *content* (vs. the *type*) of the referring expression, as when full and definite NPs are used. As indicated earlier, in these cases the constraints narrowed the search to the correct referent. The remaining cases are captured with two simple rules: if there was a simultaneous gesture to an object, then that object is the referent; otherwise the referent is the currently selected object.

While our preliminary findings accord with Huls et al., we have articulated our rules in decision list form rather than a salience ordering scheme. In fact, at least part of the Huls et al. analysis appears to be of the decision list variety, albeit cast in a salience ordering format. For instance, they found, as did we, that all referring expressions articulated with simultaneous gesturing to an object refer to that object. While they encode this preference with a very large weight (30), this value is chosen only to make certain that no other antecedent can surpass it.

To conclude, the question of whether a unified view of salience and reference for multimodal systems can be provided remains open. It appears that the nature of the tasks used in our experiments and by Huls et al. makes for a relatively easy resolution task. This could be due to two reasons: either reference is generally so constrained in multimodal interactions that the distinctions made by different referring expressions

Form	No Gesture			Simultaneous Gesture			Total
	Unselected	Selected Immediate	Selected Not Immediate	Unselected	Selected Immediate	Selected Not Immediate	
Full NP	20	32	5	10	4	0	71
Definite NP	2	1	1	0	0	0	4
"here"	0	0	0	5	3	0	8
"there"	0	7	3	0	3	1	14
"this" NP	0	0	0	2	10	0	12
"that" NP	0	1	0	0	0	0	1
"this"	0	4	0	8	5	0	17
"they"	0	1	0	0	0	0	1
"it"	0	6	0	0	2	0	8
ϕ	0	22	2	13	1	0	38
TOTAL	22	74	11	38	28	1	174

Table 1: Distribution of Referring Expressions

become unimportant for understanding, or the systems that have been developed have not been complex enough to evoke the full power of human language and gestural communication. We expect that in fact the latter is the case, and are currently designing systems in more complicated domains to test this hypothesis.

Conclusions and Future Work

We have described an implemented multimodal travel guide application being used in a WOZ setting to gather data on how successful reference is accomplished. We presented a preliminary analysis of data which suggests that, as is evident in Huls et al.'s (1995) more extensive study, the interpretation of referring expressions can be accounted for by a set of rules which do not make reference to the type of expression used. This is contrary to previous research on linguistic reference, in which the differences between such forms have been demonstrated to be crucial for understanding.

We suspect that this not a general result, but instead a product of the simplicity of the tasks around which these multimodal systems have been developed. We are currently planning the development of a crisis management scenario which would involve expert or trainee fire-fighters directing resources to objectives while using a multimodal computerized terrain model. This model will be three-dimensional and dynamic, in contrast to the two-dimensional, static map application. We expect that the complexity of the task will evoke much richer interactions, and thus may serve to clarify the use of reference in these settings.

Acknowledgements

This work was supported by National Science Foundation Grant IIS-9619126, "Multimodal Access to Spatial Data", funded within the Speech, Text, Image, and MULtimedia Advanced Technology Effort (STIMULATE).

References

Alshawi, H. 1987. *Memory and Context for Language Interpretation.* Cambridge University Press.

Cheyer, A., and Julia, L. 1995. Multimodal maps: An agent-based approach. In *Proceedings of CMC95.* 103–113.

Cheyer, A.; Julia, L.; and Martin, J.-C. 1998. A unified framework for constructing multimodal experiments and applications. In *Proceedings of CMC98*, 63–69.

Cohen, P.; Cheyer, A.; Wang, M.; and Baeg, S. 1994. An open agent architecture. In *AAAI Spring Symposium.* 1–8.

Grosz, B., and Sidner, C. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics* 12(3):175–204.

Gundel, J. K.; Hedberg, N.; and Zacharski, R. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2):274–307.

Huls, C.; Bos, E.; and Classen, W. 1995. Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics* 21(1):59–79.

Julia, L., and Cheyer, A. 1997. Speech: a privileged modality. In *Proceedings of EUROSPEECH'97.* 103–113.

Martin, J.-C. 1997. Towards intelligent cooperation between modalities. The example of a system enabling multimodal interaction with a map. In *Proceedings of the IJCAI-97 Workshop on Intelligent Multimodal Systems*. 63–69.

Martin, J.-C.; Julia, L.; and Cheyer, A. 1998. A theoretical framework for multimodal user studies. In *Proceedings of CMC98*, 104–110.

Oviatt, S. 1996. Multimodal interfaces for dynamic interactive maps. In *Proceedings of CHI96*. 95–105.

Oviatt, S.; DeAngeli, A.; and Kuhn, K. 1997. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of CHI97*. 415–422.

Prince, E. 1981. Toward a taxonomy of given-new information. In Cole, P., ed., *Radical Pragmatics*. New York, New York: Academic Press. 223–255.