In the Electronic Proceedings of the Workshop on Perceptive User Interfaces, 2001.

Using eye movements to determine referents in a spoken dialogue system

Ellen Campana
Department of Brain and Cognitive Sciences
Meliora Hall, RC Box 270268
University of Rochester
Rochester, NY 14627
1-716-275-3075
ecampana@bcs.rochester.edu

Jason Baldridge
ICCS, Division of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
44-131-650-4419
jmb@cogsci.ed.ac.uk

John Dowding
Research Institute for Advanced Computer Science
Mail Stop T27-A, NASA Ames Research Center
Moffett Field, CA 94035
1-650-604-0493
jdowding@riacs.edu

Beth Ann Hockey
Research Institute for Advanced Computer Science
Mail Stop T27-A, NASA Ames Research Center
Moffett Field, CA 94035
1-650-604-0198
bahockey@riacs.edu

Roger W. Remington
Human Factors Research and Technology Division
MS 262-4, NASA Ames Research Center
Moffett Field, CA 94035
1-650-604-6243
rremington@mail.arc.nasa.gov

Leland S. Stone
Human Factors Research and Technology Division
MS 262-2, NASA Ames Research Center
Moffett Field, CA 94035
1-650-604-3240
lstone@mail.arc.nasa.gov

ABSTRACT

Most computational spoken dialogue systems take a "literary" approach to reference resolution. With this type of approach, entities that are mentioned by a human interactor are unified with elements in the world state based on the same principles that guide the process during text interpretation. In human-to-human interaction, however, referring is a much more collaborative process. Participants often under-specify their referents, relying on their discourse partners for feedback if more information is needed to uniquely identify a particular referent. By monitoring eye-movements during this interaction, it is possible to improve the performance of a spoken dialogue system on referring expressions that are underspecified according to the literary model. This paper describes a system currently under development that employs such a strategy.

Categories and Subject Descriptors

H.5.2 [**Information systems**]: Information interfaces and presentation – *Input devices and strategies.*

General Terms

Design, Human Factors.

Keywords

Reference resolution, eye tracking, HCI, dialogue systems

1. INTRODUCTION

Most computational spoken dialogue systems take a "literary" approach to reference resolution. With this type of approach, entities that are mentioned by a human interactor are unified with elements in the world state based on the same principles that guide the process during text interpretation[2]. In human-to-human interaction, however, referring is a much more collaborative process. Participants often under-specify their referents, relying on their discourse partners for feedback if more information is needed to uniquely identify a particular referent [3,4]. This has the effect of minimizing the joint effort of the participants in an interaction by reducing the time speakers must spend specifying referents. Until recently, few computational systems have attempted to model this aspect of human communication. The difficulty lies in the fact that the cues that human interactors rely on in order to use this strategy are largely extra-linguistic, including gaze, gesture, pauses, and non-lexical grounding utterances such as "mm-hmm." Such cues occur in parallel with unfolding utterances and are used to modify utterances on-the-fly.

On the understanding end, a major difficulty has been devising a suitable way to monitor the human interactor's understanding of an utterance that the system has produced to discover when more information is needed. Generally, only verbal grounding has been considered. Systems that have speech recognition that is

PUI '01, November 15-16, 2001, Orlando, FL.

manually invoked by the user (push to talk or some variant) tend to discourage verbal grounding, or when they include it, the resulting conversation is unnatural. Even systems with continuous recognition ("open mic") tend not to utilize most grounding utterances because too little is known about how to use the varied information that is coming in. In the last few years, technological developments have provided a way to overcome these difficulties – by using eye-tracking data to monitor the human interactors' comprehension and production of referring expressions. Gaze can be used both to gain information about what a human interactor's speech refers to, and to gain information about whether a human interactor has understood language that was produced by the system.

2. GAZE AND LANGUAGE

The eye-tracking methodology has revolutionized the field of Psycholinguistics in recent years because it offers a way to monitor language comprehension as it unfolds. Participants in experiments wear a head-mounted eye-tracker, which takes images of their eyes and the scene they are viewing. This information is combined, yielding the coordinates of their gaze (eye position in world coordinates) on the computer screen they are viewing. The participants view a set of images while listening to prerecorded utterances. Researchers have investigated how language guides their attention, and therefore their eyes, to different objects on the screen as they listen to the utterances [1,15]. In another similar type of experiment, participants are asked to view objects on the screen, and then talk about those objects while their eye movements are being recorded. Researchers examined how their attention, as evidenced by eye movement, shifts from object to object while they are preparing and producing utterances [8]. The findings from such experiments suggest that input from an eye-tracker can be used for resolving underspecified referents in computational spoken dialogue systems.

On the language production end, in one experiment participants were asked to describe extemporaneously actions being carried out in individual cartoon frames. The cartoons were chosen such that some produced active descriptions and others produced passive descriptions. Regardless of syntactic type (active vs. passive voice), the data reveal that participants tended to fixate on a given entity in the cartoon roughly 900ms before the onset of the portion of their speech that referred to that entity[8].

On the language comprehension end, in one experiment participants were asked to move images around on a computer display using a mouse. The images were chosen to be related to each other phonologically to varying degrees, and all of the instructions were of the form "Pick up the X. Now put it below the Y." Language driven eye movements were observed as early as 200 ms after the onset of a target noun within a sentence [1,15].

Another language comprehension experiment looked at referring expressions, specifically. Participants in this experiment heard sentences such as "Put the apple on the towel in the box." For a given trial some participants heard this instruction while viewing a real-world scene in which there was only one apple and it was sitting on a towel. The display also contained an empty towel, an empty box, and an unrelated item. These participants were likely to make eye-movements to the empty towel just after hearing "towel." In contrast, other participants who heard the instruction while viewing a display in which there were two apples, one on a towel and one on a paper napkin, made fewer eye movements to the empty towel after hearing "the towel" in the instruction [14]. This seems to suggest that people are sensitive to the degree of specification of referring expressions, given their visual environment, from the very earliest stages of language comprehension. This supports observations that humans interpret sentences in an almost entirely incremental fashion[5].

3. DIALOGUE SYSTEM: PSA 2001

3.1 Overview

Our goal is to integrate eye-tracking information into the reference resolution portion of an existing computational spoken language dialogue interface to a simulated version of the Personal Satellite Assistant (PSA) [11]. The real PSA is a miniature robot currently being developed at NASA Ames Research Center, which is intended for deployment on the Space Shuttle and/or International Space Station (Figure 1).

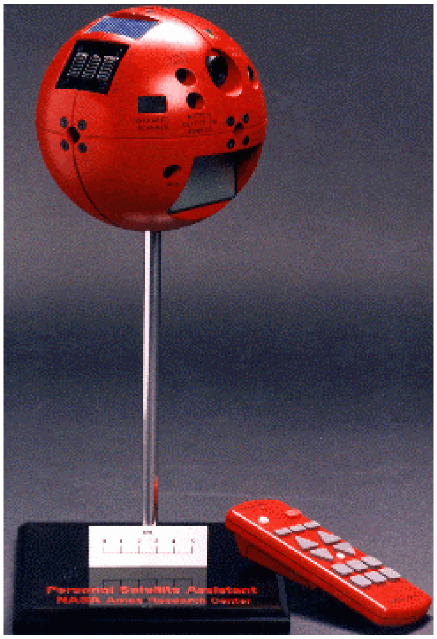

Figure 1~~1~~: Model of the PSA

It will be capable of free navigation in an indoor micro-gravity environment and will provide mobile sensory capacity as a backup to a network of fixed sensors. The PSA will primarily be controlled by voice commands through a hand-held or head-mounted microphone, with speech and language processing being handled by an offboard processor. Since the speech processing units are not in fact physically connected to the PSA we envisage that they could also be used to control or monitor other environmental functions. For example, our simulation allows voice access to the current and past values of the fixed sensor readings.

The initial PSA speech interface demo consists of a 2D simulation of the shuttle (Figure 2). State parameters include the PSA's current position, some environmental variables such as local temperature, pressure, and carbon dioxide levels, and the status of the shuttle's doors (open/closed). A visual display gives direct feedback on some of those parameters. It is this monitor that the user will be looking at when wearing the head-mounted eye-tracker[1].

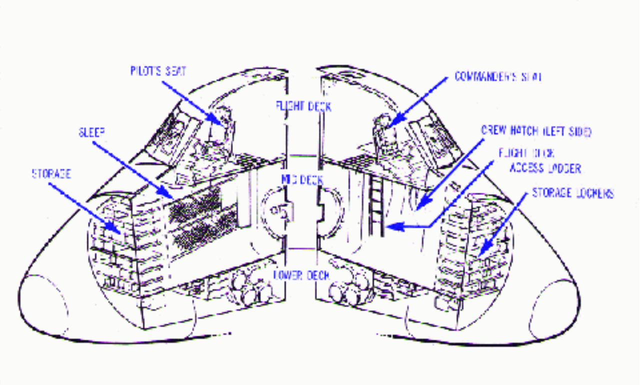

Figure 2~~2~~: PSA Simulation

The speech and language processing architecture is based on that of the SRI CommandTalk system [10,14]. The system comprises a suite of about 20 agents, connected together using the SRI Open Agent Architecture [9]. Speech recognition is performed using a version of the Nuance recognizer[12]. Initial language processing is carried out using the SRI Gemini system [6], using a domain-independent unification grammar and a domain-specific lexicon. The language processing grammar is compiled into a recognition grammar using the methods of [7]; the net result is that only grammatically well-formed utterances can be recognized. Output from the initial language-processing step is represented in a version of Quasi Logical Form [16], and passed in that form to the dialogue manager. We refer to these as *linguistic level* representations.

3.2 Reference resolution in PSA 2001

Once a linguistic level representation has been produced, the following processing steps occur:

STEP 1: The linguistic level representation is converted into a discourse level representation. This primarily involves regularizing differences in surface form: so, for example, "measure the pressure" and "what is the pressure?" have different representations at the linguistic level, but the same representation at the discourse level.

STEP 2: If necessary, the system attempts to resolve instances of ellipsis and anaphoric reference. For example, if the previous command was "measure the temperature at flight deck," then the new command "lower deck" will be resolved to an expression meaning "measure the temperature at lower deck." Similarly, if the previous command was "move to the crew hatch," then the command "open it" will be resolved to "open the crew hatch". We call the output of this step a resolved discourse level representation.

In the event of an underspecified reference, such as "open that door", if the referent of the noun phrase is not uniquely identifiable given ellipsis resolution the system replies "what door do you mean?" In some situations this would be a natural response, but seems unnatural when the interactor is focusing his/her attention on the entity that he/she is trying to refer to. It is this behavior that the system under development is designed to improve.

3.3 Improvements to PSA 2001

The physical environment of the PSA 2001 system has been augmented with an SMI Eyelink [13] head-mounted eye-tracker for monitoring eye-movements as people interact with the dialogue system. The tracker has a temporal resolution of 4 msec and an average error of in gaze-position accuracy on the order of .5°-1.0°.

The dialogue system itself has been improved by the addition of a multi-modal interface (MMI) software agent. This agent was developed using Java interfaces to maximize extensibility for future applications. The MMI gives us the ability to separate portions of the system that implement distinct functionality and to develop heuristics to choose between their potentially contradictory outputs. It provides a framework and much of the functionality for handling input from the eye tracker, and in the future from additional input devices such as a glove, light pen, joystick or mouse. It will also allow for integration of input information across modalities and across multiple instances of the same modality, for instance several eye-trackers or several event-handlers corresponding to the same eye-tracker.

We have added eye-tracking capabilities to the PSA 2001 in order to improve reference resolution in the case where there are multiple possible referents to a noun phrase uttered by the user, and the noun phrase is underspecified according to the literary model of reference resolution. Since people tend to fixate on the objects they mention 900 ms before beginning to mention them, regardless of syntactic type, gaze can be used in step two of the process described above. If the ellipsis resolution fails to identify the referent of an anaphor, eye fixations in the second preceding the production of the anaphor will be used to disambiguate it. Thus, if the user looks at the crew hatch just before saying "door" in the command "open that door" then the resolved discourse level representation will be something representing "open the crew hatch." Similarly, if the user looks at the mid deck just before saying "there" in the command "go there" then the resolved discourse level representation will be something representing "go to mid deck." We expect to see a reduction in the number of turns required to complete tasks in the PSA environment, as well as a reduction in overall time, as the strategy we are implementing minimizes joint effort.

In cases where the interactor uses "the" rather than "that" in the command, eye-tracking will be used to disambiguate unidentified referents but only if the PSA is not in close proximity to a potential referent. For instance, if the PSA is physically located next to the crew hatch door and the interactor says "open the door" the resolved discourse level representation will be something representing "open the crew hatch," regardless of where the interactor's eyes are fixated.

Eye tracking in the simulation is also used to address spoken commands to either the simulated PSA or the simulated astronaut. By fixating on one of these entities, speech recognition is activated and spoken input is interpreted as a command to that agent. In actual deployment in the ISS this capability translates into the ability to address one PSA out of several visually identical ones that would be deployed.

[1] Floor-mounted eye-trackers could also be used in the future, as technology improves.

4. CONCLUDING REMARKS

In this paper we have described a set of modifications to the PSA 2001 system that will use eye-movement data to improve the performance of the dialogue system. We realize that there are a myriad of other interesting applications of eye-tracking within an asynchronous agent-based spoken dialogue system. This is an ongoing project and we plan to investigate a number of other avenues for development. For instance, in a future version of the system, eye movements will be considered during incremental production of referring expressions. The PSA mobile unit is equipped with sensors that allow it to detect problems on the shuttle, such as fires or pockets of dangerous fumes. It also monitors fixed sensors in remote areas of the shuttle. In the event that it finds such a hazard, it must notify the human interactors. Currently, it produces a sentence containing a fully specified referent, for instance "I have detected high levels of carbon dioxide on mid deck near the storage lockers." Eye-movement data will allow it to say "I have detected high levels of carbon dioxide near the storage lockers" and add "on mid deck" only if the interactor's eyes indicate difficulty in disambiguating the referent of "storage lockers." Again this is a shift toward more natural interaction, which will produce gains in terms of reduced time to complete tasks in the PSA demo environment.

As we have argued, eye movement data is useful for the portions of the dialogue system that deal with reference resolution, but we also plan to research its utility in other portions of the system. Because we have chosen to implement this capability within an asynchronous, agent-based framework, the eye-movement data will be available to the speech recognizer and the planner. In the future we hope to implement a version of the system in which eye movement data is used to boost recognition in noisy environments by providing confidence scores associated with individual words at the time of recognition. We also hope to use these confidence scores to select appropriate confirmation strategies. For instance, if the recognizer outputs "go to mid deck" for a given portion of speech, and the interactor was looking at mid deck in the 2D simulation 900ms before mentioning it, no confirmation is needed. If, on the other hand, the interactor was looking at flight deck at that time, a better confirmation strategy would be for the PSA to reply with either, " I am proceeding to mid deck" as it begins to move (confirmation without waiting), or "I will proceed to mid deck, Okay?", waiting for the user's response before executing the command (confirm before execution). At present all three confirmation strategies are available in the system one at a time by setting switches. Having both eyetracker data and confidence scores will provide a better basis for switching between them in a principled way.

The MMI Agent has been designed in a general way so that it will be able to handle input from many different modalities in addition to eye-movement data, such as data gloves, light pens, mice, and joysticks. Multiple instances of the same input device could also be integrated using the MMI. Using these multiple sources of input to improve the dialogue system will be the subject of future developments to the system.

REFERENCES

[1] Allopenna, P., Magnuson, J., and Tanenhaus, M., 1998. Tracking the time course of spoken word recognition. *Journal of Memory and Language*, vol. 38, pages 419-439.

[2] Baldwin, B., 1995. CogNiac: A discourse processing engine. Ph.D. Thesis, University of Pennsylvania, Department of Computer and Information Sciences.

[3] Clark, H., and Schaefer, E., 1989. Collaborating on contributions to converstations. In Dietrich, R., and Graumann, C. (eds.) *Language processing in Social Contexts*. Elsevier Press.

[4] Clark, H., and Wilkes-Gibbs, D., 1986. Referring as a collaborative process. *Cognition*, vol. 22, pages 1-39.

[5] Crain, S. and Steedman, M., 1985. On not being led up the garden path: the use of context by the Psychological Parser. In Dowty, D., Kartunnen, L., and Zwicky, A. (eds.) *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives.* Cambridge University Press.

[6] Dowding, J., Gawron, M., Appelt, D., Cherny, L., Moore, R. and Moran, D. 1993. Gemini: A natural language system for spoken language understanding. In *Proceedings for the Thirty-First Annual Meeting of the Association for Computational Linguistics*.

[7] Dowding, J., Hockey, B. A., Gawron, M. J., and Culy, C. 2001. Practical issues in compiling typed unification grammars for speech recognition. Proceedings for the 39th Annual Meeting of the Association for Computational Linguistics. Toulouse, France.

[8] Griffin, Z., and Bock, K. 2000. What the eyes say about speaking. *Psychological Science*, vol. 11(4), pages 274-279.

[9] Martin, D., Cheyer, A. and Moran, D. 1998. Building distributed software systems with the open agent architecture. In *Proceedings of the Third International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology*.

[10] Moore, R., Dowding, J., Bratt, H., Gawron, J., Gorfu, Y., and Cheyer, A. 1997. CommandTalk: A spoken-language interface for battlefield simulations. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 1-7.

[11] PSA, 2001. *Personal Satellite Assistant (PSA) Project*. http://ic.arc.nasa.gov/ic/psa. As of 2 July 2001.

[12] Nuance, 2001. Nuance Communications, Inc. http://www.nuance.com. As of July 2, 2001.

[13] SMI, 2001. Sensomotoric Instruments, Inc. http://www.smi.de/el/index.html. As of 2 July 2001.

[14] Stent, A., Dowding, J., Gawron, J., Bratt, E., and Moore, R. 1999. The CommandTalk spoken dialog system. In *Proceedings of the Thirty-Seventh Annual Meeting of the Association for Computational Linguistics*, pages 183-190.

[15] Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., and Sedivy, J., 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, vol. 268, pages 632-634.

[16] Van Eijck, J., and Moore, R. 1992. Semantic rules for English. In Alshawi, H. (editor) *The Core Language Engine*. MIT Press