Our Dinner with Cassie

Stuart C. Shapiro and **Haythem O. Ismail** and **John F. Santore**
Department of Computer Science and Engineering
and Center for Cognitive Science
University at Buffalo
226 Bell Hall
Buffalo, NY 14260-2000
{shapiro}{hismail}{jsantore}@cse.buffalo.edu

Abstract

This paper summarizes over a decade of research on natural language dialogue with robotic agents, and their required underlying knowledge representation and reasoning abilities. The major features that characterize robots as spoken dialogue partners with humans, are: 1) the robot and human share a perceptual field containing objects they can both point to and manipulate; 2) robots are actors as well as speakers and hearers.

Introduction

Spoken natural language is the medium of the future for human interaction with machines, because it is the easiest, most natural modality for people to use. This is becoming widely evident as speech has become commercially viable for dictation to word processors, input to telephonic menus, and for command of desktop computers, VCRs, car phones, etc. However, virtually all current such systems (other than the dictation systems) use speech as nothing much more than a simple command language. More robust understanding is needed to make the interaction natural for people, and this depends on the current and future research of the people who are present at this Symposium, among others.

The major features that characterize robots as spoken dialogue partners with humans, are: 1) the robot and human share a perceptual field containing objects they can both point to and manipulate; 2) robots are actors as well as speakers and hearers.

The senior author of this paper (Shapiro) has been involved in research on natural language (including speech) dialogue with robotic agents, and the necessary underlying knowledge representation and reasoning (KRR) abilities, for over a decade, as part of the ongoing Cassie projects of NL-competent computational agents (Shapiro & Rapaport 1987; Shapiro 1989; Shapiro & Rapaport 1991; Shapiro 1998).

CUBRICON

The CUBRICON project (Neal *et al.* 1988; 1989a; 1989b; 1998; Neal & Shapiro 1991; 1994), developed a prototype intelligent multimodal interface between a human and an air mission planning system. The computer displays, which comprised the environment shared between the human user and the interface agent, consisted of one screen containing various windows showing maps, and one screen containing textual forms. On the maps were icons representing places, areas such as airbases, and objects such as missle batteries. User input was in the form of typed text, speech, and one mouse button for pointing. Typed text and speech could be in a fragment of English. The mouse could be used to point to windows, maps, map icons, and form fields. Computer output was in the form of speech, text, and creation and/or rearrangement of maps and windows. The interface agent could point by blinking window frames, blinking icons, drawing circles around icons, drawing arrows to icons, and drawing arrows from one icon to another to express spatial relations.

Input streams were merged, and the grammar recognized multi-modal references. For example, the grammar considered a noun phrase to be an appropriate string of words (as usual) possibly augmented by initial, internal, and/or final pointing gestures. Multiple noun pharses in a sentence could each contain their own set of pointing gestures. Multi-modal noun phrases were interpreted as referring to entities that satisfied the properties mentioned in the linguistic portion of the noun phrase, while being represented by icons in the vicinity of the pointing gesture. The linguistic portion of the input was used to disambiguate whether the mouse pointing was to a map icon, the map on which the icon was, or the window containing the map.

Output pointing gestures were timed to occur just before, during, or just after the linguistic part of the noun phrase they were a part of. Multiple output noun phrases could each contain their own set of pointing gestures.

This work made it clear that referring expressions using a combination of NL and pointing are less ambiguous than either one alone. This lesson is directly applicable to natural communication systems between humans and practical robots.

Discussing and Using Plans

A project on "Discussing and Using Plans" (Kumar, Ali, & Shapiro 1988; Shapiro *et al.* 1989; 1990; Shapiro, Kumar, & Ali 1990) involved the development of KRR techniques for representing acts and plans so that they could be discussed in NL and reasoned about, as well as performed. This project culminated in the SNeRE BDI model (Kumar 1990; 1993; 1996; Kumar & Meeden 1998; Kumar & Shapiro 1991; 1993; 1994a; 1994b) that allows for reasoning in the service of acting, and acting in the service of reasoning, as well as performing, discussing, and reasoning about acts and plans. A simulated blocks-world robot developed during this project could understand NL explanations of how to perform blocks-world activities, and then reason about and discuss the instructions, and also behave according to them. For example, the following is an extract of the instructions the blocks-world robot could understand and follow.

> *There is a table. The table is a support. Blocks are supports. Before picking up a block the block must be clear. Before putting a block on a support the support must be clear. After putting a block on another block the latter block is not clear. If a block is on a support then a plan to achieve that the support is clear is to pick up the block and then put the block on the table. A plan to pile a block on another block on a third block is to put the third block on the table and then put the second block on the third block and then put the first block on the second block.*

Natural communication with practical robots will have to include this ability for the human to explain to the robot what the robot is to do.

GLAIR

The GLAIR (Grounded Layered Architecture with Integrated Reasoning) agent architecture (Hexmoor, Lammens, & Shapiro 1993; Hexmoor *et al.* 1993; Lammens, Hexmoor, & Shapiro 1995; Hexmoor 1995; Hexmoor & Shapiro 1997) has been developed for robots and other agents that use NL and various sensors and effectors. GLAIR is a three-level architecture consisting of:

The Knowledge Level (KL): the location of symbolic "conscious" reasoning, implemented by the SNePS Knowledge Representation and Reasoning system (Shapiro & Rapaport 1992; Shapiro & The SNePS Implementation Group 1999), in which terms of the SNePS logical language represent the mental entities conceived of and reasoned about by Cassie, the robotic agent;

The Perceptuo-Motor Level (PML): the location of routine behaviors that can be carried out without thinking about each step, and of the data objects that these behaviors operate on;

The Sensori-Actuator Level (SAL): the location of control of individual sensors and effectors.

A major theme of GLAIR is the alignment (a variety of symbol-grounding) of KRR terms denoting objects and acts with their corresponding sensory/effector representations to tighten the cross-modal correlation of language, sensing, and acting. KRR object-denoting or category-denoting terms are aligned with symbols, constituting descriptions, that the vision system (real or simulated) can use to locate the corresponding objects in the (real or simulated) world. Also these symbols, passed up from the SAL to the PML, can be used to recognize the KR terms that denote entities with those descriptions. KRR action-denoting terms are aligned with PML behaviors that carry out the actions.

FEVAHR

The FEVAHR project (Shapiro 1998) uses the GLAIR architecture to have Cassie play the role of a "Foveal Extra-Vehicular Activity Helper-Retriever (FEVAHR)." Cassie, the FEVAHR, was implemented on a commercial Nomad robot, including sonar, bumpers, and wheels, enhanced with a foveal vision system consisting of a pair of cameras with associated hardware and software. There have also been several software simulatated versions of the FEVAHR. FEVAHR/Cassie operates in a $17' \times 17'$ room containing: Cassie; Stu, a human supervisor; Bill, another human; a green robot; and three indistinguishable red robots. Cassie is always talking to either Stu or Bill—taking statements, questions, and commands from that person (all expressed in a fragment of English), and responding and reporting to that person in English. Cassie can be told, by the person addressing her, to talk to the other person, or to find, look at, go to, or follow any of the people or robots in the room. Cassie can also engage in conversations on a limited number of other topics in a fragment of English, similar to some of the conversations in (Shapiro 1989).

FEVAHR/Cassie is able to understand and use deictic terms including "I," "you," "come," and "go," by maintaining a representation of itself, and a belief about whom it is conversing with. It maintains a concept of personal time, including a NOW pointer which is updated whenever it acts, and which is used to determine the tense and aspect of sentences it utters to report on its current and past actions. A summary of FEVAHR/Cassie's use of deixis is,

- Cassie's interpretation of "here" and "I" depend on who is talking to her.
- Cassie addresses whomever she is talking to as "you".
- Cassie understands that when Stu or Bill use "you" they mean her, and she has beliefs about herself which she expresses using "I".

- Cassie uses present tense when reporting her current actions, but past tense to report past actions, even though those actions were reported in the present tense while they were occurring.
- The interpretation of indexicals is done by the analysis grammar; the generation of indexicals is done by the generation grammar; the KR representation remains the same.

Below is a transcript of an interaction with FEVAHR/Cassie. When it begins, Cassie is looking at and talking to Stu, but not near him. Stu's and Bill's utterances are in *italics*. Cassie's utterances are in `typewriter` font.

Come here.
`I came to you, Stu.`
`I am near you.`

The green robot is Robbie.
`I understand that Robbie is the green robot.`

Go to the green robot.
`I found Robbie.`
`I am looking at Robbie.`
`I went to Robbie.`
`I am near Robbie.`

Talk to Bill.
`I am talking to you, Bill.`

Come here.
`I found you.`
`I am looking at you.`
`I came to you.`
`I am near you.`

Where were you?
`I was near Stu`
`and I was near Robbie`
`and I am near you.`

Who have you talked to?
`I talked to Stu`
`and I am talking to you.`

You are smart.
`I understand that I am smart.`

The software simulated versions of FEVAHR/Cassie can understand NPs with associated pointing gestures, although not in as sophisticated a fashion as CUBRICON. The vision folks never implemented gaze control for pointer-following in the hardware version.

UXO Remediation

The UXO Remediation project (Ismail & Shapiro 2000) is a modification of the FEVAHR robot to implement NL-competent robots that clean up unexploded ordnance (UXO remediation). This Cassie has only existed as a software simulation. The UXO-clearing-Cassie exists in an area consisting of four zones: a safe zone; an operating zone that possibly contains UXOs; a drop-off zone; and a recharging zone. The UXO-clearing-Cassie contains a battery that discharges as she operates, and must be recharged in the recharge zone as soon as it reaches a low enough level. She may carry charges to use to blow up UXOs. Her task is to search the operating zone for a UXO, and either blow it up by placing a charge on it, and then going to a safe place to wait for the explosion, or pick up the UXO, take it to the drop-off zone, and leave it there. The UXO-clearing-Cassie has to interrupt what she is doing whenever the battery goes low, and any of her actions might fail. (She might drop a UXO she is trying to pick up.) She takes direction from a human operator in a fragment of English, and responds and reports to that operator. There is a large overlap in the grammars of FEVAHR/Cassie and the UXO-clearing-Cassie.

There are two main active areas of research using the UXO-clearing-Cassie:

1. One area involves issues of the representation of and reasoning about NOW by a reasoning, acting, natural language competent agent. We have been working on solutions of two problems in this area (Ismail & Shapiro 2000):
   (a) The problem of "the unmentionable now" results from the inability to refer to future values of the variable NOW. Since NOW can only refer to the current time of an assertion or action (mirroring the behavior of the English "now"), one cannot use it in the KRR object language to refer to the future. Such reference to future now's is important for specifying conditional acts and acting rules. Our solution is to eliminate any reference to those times in the object language, but to modify the forward and backward chaining procedures so that they insert the appropriate values of NOW at the time of performing a conditional act or using an acting rule.
   (b) The problem of "the fleeting now" emerges when, in the course of reasoning about (the value of) NOW, the reasoning process itself results in NOW changing. Our solution is based on realizing that, at any point, the value of NOW is not a single term, but rather a stack of terms. Each term in the stack corresponds to the agent's notion of the current time at a certain level of granularity, with granularity growing coarser towards the bottom of the stack. Temporal progression and granularity shifts are modeled by various stack operations.
2. A second area of current research concerns understanding NPs whose referents can only be determined after sensory and other actions. (*See also* (Haas 1995).) In some cases, the referent is an entity the robot can recognize, such as when FEVAHR/Cassie

is told

Go to a person.

In some cases, even though the referent looks like a previously encountered entity, it is new, as when the UXO-clearing-Cassie is told

Find a UXO.

In other cases, the referent may be completely new, such as

Clean up the field from the tree over there to the creek that's just beyond that hill.

Conclusions

Spoken natural language is the medium of the future for human interaction with practical robots. However, more research is needed on robust understanding to make the interaction natural. The major features that characterize robots as spoken dialogue partners with humans, are: 1) the robot and human share a perceptual field containing objects they can both point to and manipulate; 2) robots are actors as well as speakers and hearers. Naturally communicating practical robots must include: the ability to understand and use language combined with pointing gestures; the ability to understand instructions given in natural language, and then to behave according to those instructions; the correlation of language, reasoning, sensing, and acting, and the terms and symbols used by those modalities; the ability to report on what they are doing, and to remember and report on what they have done; a personal sense of time, and a correct use of temporal references; the ability to use and understand language referring to the not-here and the not-now, as well as to the here and now.

Acknowledgments

The research reported in this paper was conducted by the authors and other members of the SNePS Research Group, Department of Computer Science and Engineering (previously the Department of Computer Science), State University of New York at Buffalo. The research was supported as follows.

CUBRICON: supported in part by the Defense Advanced Research Projects Agency and monitored by the Rome Air Development Center under contract no. F30603-87-C-0136.

Discussing and Using Plans: supported in part by the Air Force Systems Command, Rome Air Development Center, Griffiss Air Force Base, New York 13441-5700, and the Air Force Office of Scientific Research, Bolling AFB DC 20332 under Contract No. F30602-85-C-0008, which supported the Northeast Artificial Intelligence Consortium (NAIC).

GLAIR: supported in part by Equipment Grant No. EDUD-US-932022 from SUN Microsystems Computer Corporation.

FEVAHR: supported in part by NASA under contracts NAS 9-19004 and NAS 9-19335.

UXO Remiadiation: supported in part by ONR under contract N00014-98-C-0062.

References

Haas, A. R. 1995. Planning to find the referents of noun phrases. *Computational Intelligence* 11(4):593–624.

Hexmoor, H., and Shapiro, S. C. 1997. Integrating skill and knowledge in expert agents. In Feltovich, P. J.; Ford, K. M.; and Hoffman, R. R., eds., *Expertise in Context*. Cambridge, MA: AAAI Press/MIT Press. 383–404.

Hexmoor, H.; Lammens, J.; Caicedo, G.; and Shapiro, S. C. 1993. Behaviour based AI, cognitive processes, and emergent behaviors in autonomous agents. In Rzevski, G.; Pastor, J.; and Adey, R., eds., *Applications of Artificial Intelligence in Engineering VIII*, volume 2. Southampton and London: Computational Mechanics Publications with Elsevier Applied Science. 447–461.

Hexmoor, H.; Lammens, J.; and Shapiro, S. C. 1993. Embodiment in GLAIR: a grounded layered architecture with integrated reasoning for autonomous agents. In Dankel II, D. D., and Stewman, J., eds., *Proceedings of The Sixth Florida AI Research Symposium (FLAIRS 93)*. The Florida AI Research Society. 325–329.

Hexmoor, H. H. 1995. *Representing and Learning Routine Activities*. Ph.D. Dissertation, Department of Computer Science, State University of New York at Buffalo, Buffalo, NY. Technical Report 98-04.

Ismail, H. O., and Shapiro, S. C. 2000. Two problems with reasoning and acting in time. In Cohn, A. G.; Giunchiglia, F.; and Selman, B., eds., *Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference (KR 2000)*. San Francisco: Morgan Kaufmann. in press.

Kumar, D., and Meeden, L. 1998. A hybrid connectionist and BDI architecture for modeling enbedded rational agents. In *Cognitive Robotics: Papers from the 1998 AAAI Fall Symposium, Technical Report FS-98-02*. Menlo Park, California: AAAI Press. 84–90.

Kumar, D., and Shapiro, S. C. 1991. Modeling a rational cognitive agent in SNePS. In Barahona, P.; Pereira, L. M.; and Porto, A., eds., *EPIA 91: 5th Portugese Conference on Artificial Intelligence, Lecture Notes in Artificial Intelligence 541*. Heidelberg: Springer-Verlag. 120–134.

Kumar, D., and Shapiro, S. C. 1993. Deductive efficiency, belief revision and acting. *Journal of Experimental and Theoretical Artificial Intelligence (JETAI)* 5(2&3):167–177.

Kumar, D., and Shapiro, S. C. 1994a. Acting in service of inference (and *vice versa*). In Dankel II, D. D., ed., *Proceedings of The Seventh Florida AI Research Symposium (FLAIRS 94)*. The Florida AI Research Society. 207–211.

Kumar, D., and Shapiro, S. C. 1994b. The OK BDI architecture. *International Journal on Artificial Intelligence Tools* 3(3):349–366.

Kumar, D.; Ali, S.; and Shapiro, S. C. 1988. Discussing, using and recognizing plans in SNePS preliminary report—SNACTor: An acting system. In Rao, P. V. S., and Sadanandan, P., eds., *Modern Trends in Information Technology: Proceedings of the Seventh Biennial Convention of South East Asia Regional Computer Confederation*. New Delhi, India: Tata McGraw-Hill. 177–182.

Kumar, D. 1990. An integrated model of acting and inference. In Kumar, D., ed., *Current Trends in SNePS*. Berlin: Springer-Verlag Lecture Notes in Artificial Intelligence, No. 437. 55–65.

Kumar, D. 1993. A unified model of acting and inference. In *Proceedings of the Twenty-Sixth Hawaii International Conference on System Sciences*. Los Alamitos, CA: IEEE Computer Society Press.

Kumar, D. 1996. The SNePS BDI architecture. *Decision Support Systems* 16:3.

Lammens, J. M.; Hexmoor, H. H.; and Shapiro, S. C. 1995. Of elephants and men. In Steels, L., ed., *The Biology and Technology of Intelligent Autonomous Agents*. Berlin: Springer-Verlag, Berlin. 312–344.

Lehmann, F., ed. 1992. *Semantic Networks in Artificial Intelligence*. Oxford: Pergamon Press.

Neal, J. G., and Shapiro, S. C. 1991. Intelligent multimedia interface technology. In Sullivan, J. W., and Tyler, S. W., eds., *Intelligent User Interfaces*. Reading, MA: Addison-Wesley. 11–43.

Neal, J. G., and Shapiro, S. C. 1994. Knowledge-based multimedia systems. In Buford, J. F. K., ed., *Multimedia Systems*. Reading, MA: ACM Press/Addison Wesley. 403–438.

Neal, J. G.; Dobes, Z.; Bettinger, K. E.; and Byoun, J. S. 1988. Multi-modal references in human-computer dialogue. In *Proceedings of the Seventh National Conference on Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann. 819–823.

Neal, J. G.; Thielman, C. Y.; Funke, D. J.; and Byoun, J. S. 1989a. Multi-modal output composition for human-computer dialogues. In *Proceedings of the 1989 AI Systems in Government Conference*, 250–257.

Neal, J. G.; Thileman, C. Y.; Dobes, Z.; Haller, S. M.; and Shapiro, S. C. 1989b. Natural language with integrated deictic and graphic gestures. In *Proceedings of the DARPA Speech and Natural Language Workshop*. Los Altos, CA: Morgan Kaufmann. 410–423.

Neal, J. G.; Thielman, C. Y.; Dobes, Z.; Haller, S. M.; and Shapiro, S. C. 1998. Natural language with integrated deictic and graphic gestures. In Maybury, M. T., and Wahlster, W., eds., *Readings in Intelligent User Interfaces*. San Francisco: Morgan Kaufmann. 38–51.

Shapiro, S. C., and Rapaport, W. J. 1987. SNePS considered as a fully intensional propositional semantic network. In Cercone, N., and McCalla, G., eds., *The Knowledge Frontier*. New York: Springer-Verlag. 263–315.

Shapiro, S. C., and Rapaport, W. J. 1991. Models and minds: Knowledge representation for natural-language competence. In Cummins, R., and Pollock, J., eds., *Philosophy and AI: Essays at the Interface*. Cambridge, MA: MIT Press. 215–259.

Shapiro, S. C., and Rapaport, W. J. 1992. The SNePS family. *Computers & Mathematics with Applications* 23(2–5):243–275. Reprinted in (Lehmann 1992, pp. 243–275).

Shapiro, S. C., and The SNePS Implementation Group. 1999. *SNePS 2.5 User's Manual*. Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY.

Shapiro, S. C.; Woolf, B.; Kumar, D.; Ali, S. S.; Sibun, P.; Forster, D.; and Anderson, S. 1989. Northeast artificial intelligence consortium annual report – 1988: Discussing, using, and recognizing plans. Report RADC-TR-89-259, Rome Air Development Center, Griffiss Air Force Base.

Shapiro, S. C.; Woolf, B.; Kumar, D.; Ali, S. S.; Sibun, P.; Forster, D.; Anderson, S.; Pustejovesky, J.; and Haas, J. 1990. Discussing, using, and recognizing plans–Project Report. Technical Report RADC-TR-90-404, Volume II (of 18), North-East Artificial Intelligence Consortium, Griffiss Air Force Base, NY.

Shapiro, S. C.; Kumar, D.; and Ali, S. 1990. A propositional network approach to plans and plan recognition. In Maier, A., ed., *Proceedings of the 1988 Workshop on Plan Recognition*. San Mateo, CA: Morgan Kaufmann.

Shapiro, S. C. 1989. The CASSIE projects: An approach to natural language competence. In Martins, J. P., and Morgado, E. M., eds., *EPIA 89: 4th Portugese Conference on Artificial Intelligence Proceedings, Lecture Notes in Artificial Intelligence 390*. Berlin: Springer-Verlag. 362–380.

Shapiro, S. C. 1998. Embodied Cassie. In *Cognitive Robotics: Papers from the 1998 AAAI Fall Symposium, Technical Report FS-98-02*. Menlo Park, California: AAAI Press. 136–143.