Cognitive Robotics: Command, Interrogation and Teaching in Robot Coaching

Alfredo Weitzenfeld[1] and Peter Ford Dominey[2]

[1] ITAM, San Angel Tizapán, México DF, CP 0100
alfredo@itam.mx
http://www.cannes.itam.mx/Alfredo/English/Alfredo.htm
[2] Institut des Sciences Cognitives, CNRS, 67 Blvd. Pinel, 69675 Bron Cedex, France
dominey@ isc.cnrs.fr
http://www.isc.cnrs.fr/dom/dommenu-en.htm

Abstract. The objective of the current research is to develop a generalized approach for human-robot interaction via spoken language that exploits recent developments in cognitive science, particularly notions of grammatical constructions as form-meaning mappings in language, and notions of shared intentions as distributed plans for interaction and collaboration. We demonstrate this approach distinguishing among three levels of human-robot interaction. The first level is that of commanding or directing the behavior of the robot. The second level is that of interrogating or requesting an explanation from the robot. The third and most advanced level is that of teaching the robot a new form of behavior. Within this context, we exploit social interaction by structuring communication around shared intentions that guide the interactions between human and robot. We explore these aspects of communication on distinct robotic platforms, the Event Perceiver and the Sony AIBO robot in the context of four-legged RoboCup soccer league. We provide a discussion on the state of advancement of this work.

1 Introduction

Ideally, research in Human-Robot Interaction will allow natural, ergonomic, and optimal communication and cooperation between humans and robotic systems. In order to make progress in this direction, we have identified two major requirements: First, we must work in real robotics environments in which technologists and researchers have already developed an extensive experience and set of needs with respect to HRI. Second, we must develop a domain independent language processing system that can be applied to arbitrary domains and that has psychological validity based on knowledge from social cognitive science. In response to the first requirement regarding the robotic context, we have studied two distinct robotic platforms. The first, the *Event Perceiver* is a system that can perceive human events acted out with objects, and can thus generate descriptions of these actions. The second is the *Sony AIBO* robot having local visual processing capabilities in addition to autonomous mobility. In the latter, we explore human-robot interaction in the context of four-legged RoboCup soccer league. From the psychologically valid language context, we base the interactions on a model of language and meaning correspondence

G. Lakemeyer et al. (Eds.): RoboCup 2006, LNAI 4434, pp. 379–386, 2007.

developed by Dominey et al. [1] having described both neurological and behavioral aspects of human language, and having been deployed in robotic contexts, and second, on the notion of shared intentions or plans by Tomasello et al. [2, 3] that will be used to guide the collaborative interaction between human and robot. The following sections describe the platforms, the spoken language interface for command, control and teaching these systems, and current experimental results with the Sony AIBO platform.

2 Cognitive Robotics: A Spoken Language Approach

In Dominey & Boucher [4, 5, 6] we describe the **Event Perceiver System** that could adaptively acquire a limited grammar based on training with human narrated video events. An image processing algorithm extracts the meaning of the narrated events translating them into *action*(*agent, object, recipient*) descriptors. The event extraction algorithm detects physical contacts between objects (see [7]), and then uses the temporal profile of contact sequences in order to categorize the events. The visual scene processing system is similar to related event extraction systems that rely on the characterization of complex physical events (e.g. give, take, stack) in terms of composition of physical primitives such as contact (e.g. [8, 9]). Together with the event extraction system, a speech to text system was used to perform translations sentence to meaning using different languages [10].

2.1 Processing Sentences with Grammatical Constructions

Each narrated event generates a well formed <*sentence*, *meaning*> pair that is used as input to a model that learns the sentence-to-meaning mappings as a form of template in which nouns and verbs can be replaced by new arguments in order to generate the corresponding new meanings. These templates or grammatical constructions (see [11]) are identified by the configuration of grammatical markers or function words within the sentences [12].

Table 1. Sentences and corresponding constructions

	Sentence	Construction <sentence, meaning>
1	The robot kicked the ball	<*Agent event object, event(agent, object*>
2	The ball was kicked by the robot	<*Object* was *event* by *agent, event(agent, object*>
3	The red robot gave the ball to the blue robot	<*Agent event object* to *recipient, event(agent, object, recipient)*>
4	The ball was given to the blue robot by the red robot	<*Object* was *event* to *recipient* by *agent, event(agent, object, recipient)*>
5	The blue robot was given the ball by the red robot	<*Recipient* was *event object* by *agent, event(agent, object, recipient)*>

Each grammatical construction corresponds to a mapping from sentence to meaning. This information is also used to perform the inverse transformation from meaning to sentence. For the initial sentence generation studies we concentrated on the 5 grammatical constructions shown in Table 1. These correspond to constructions with one verb and two or three arguments in which each of the different arguments

can take the focus position at the head of the sentence. On the left example sentences are presented, and on the right, the corresponding generic construction is shown. In the representation of the construction, the element that will be at the pragmatic focus is underlined.

This construction set provides sufficient linguistic flexibility, for example, when the system is interrogated about the red robot, the blue robot or the ball. After describing the event *give(red robot, blue robot, ball)*, the system can respond appropriately with sentences of type 3, 4 or 5, respectively. Note that sentences 1-5 are specific sentences that exemplify the 5 constructions in question, and that these constructions each generalize to an open set of corresponding sentences.

We have used the CSLU Speech Tools Rapid application Development (RAD) [13] to integrate these pieces, including (a) scene processing for event recognition, (b) sentence generation from scene description and response to questions, (c) speech recognition for posing questions, and (d) speech synthesis for responding.

2.2 Shared Intentions for Learning

Perhaps the most interesting aspect of the three part "command, interrogate, teach" scenario involves learning. Our goal is to provide a generalized platform independent learning capability that acquires new <*percept*, *response*> constructions. That is, we will use existing perceptual capabilities, and existing behavioral capabilities of the given system in order to bind these together into new, learned <*percept*, *response*> behaviors.

The idea is to create new <*percept*, *response*> pairs that can be permanently archived and used in future interactions. Ad-hoc analysis of human-human interaction during teaching-learning reveals the existence of a general intentional plan that is shared between teachers and learners, which consists of three components. The first component involves specifying the percept that will be involved in the <*percept*, *response*> construction. This percept can be either a verbal command, or an internal state of the system that can originate from vision or from another sensor. The second component involves specifying what should be done in response to this percept. Again, the response can be either a verbal response or a motor response from the existing behavioral repertoire. The third component involves the binding together of the <*percept*, *response*> construction, and validation that it was learned correctly. This requires the storage of this new construction in a construction database so that it can be accessed in the future. This will permit an open-ended capability for a variety of new types of communicative behavior.

In the following section this capability is used to teach a robot to respond with physical actions or other behavioral responses to perceived objects or changes in internal states. The user enters into a dialog context, and tells the robot that we are going to learn a new behavior. The robot asks *what is the perceptual trigger of the behavior* and the human responds. The robot then asks *what is the response behavior*, and the human responds again. The robot links the <*percept*, *response*> pair together so that it can be used in the future.

Having human users control and interrogate robots using spoken language results in the ability to ergonomically teach robots. Additionally, it is also useful to execute components of these action sequences conditional on perceptual values. For example

the user might want to tell the robot to walk forward until it comes close to an obstacle, using a "command *X* until *Y*" construction, where *X* corresponds to a continuous action (e.g. walk, turn left) and *Y* corresponds to a perceptual condition (e.g. collision detected, ball seen, etc.).

3 Human-Robot Coaching in RoboCup Soccer

In order to demonstrate the generalization of the spoken language human-robot interaction approach we have begun a series of experiments in the domain of RoboCup Soccer [14], a well documented and standardized robot environment thus provides a quantitative domain for evaluation of success. For this project we have chosen as testing platform the Four-Legged league where ITAM's Eagle Knights team regularly competes [15, 16]. In this league two teams of four robots play soccer on a small-carpeted soccer field using Sony's Four-Legged AIBO robots. While no human intervention is allowed during a game, in the future humans could play a decisive role analogous to real soccer coaches adjusting in real-time their team playing characteristics according to the state of the game, individual or group performance. While no such human interaction is possible in the Four-Legged league, RoboCup incorporates a simulated coaching league where coaching agents can learn during a game and then advice virtual soccer agents on how to optimize their behavior accordingly (see [17, 18]).

3.1 Human-Robot Architecture

The human-robot interaction architecture is illustrated in Figure 2. The spoken language interface is provided by the CSLU-RAD framework while communication to the Sony AIBO robots is done in a wireless fashion via the CMU Tekkotsu platform [19] and URBI [21]. The CMU Tekkotsu and URBI systems provide a high level interface for remotely controlling the AIBO. Via this interface, the AIBO can be commanded to perform different actions as well as be interrogated with respect to various internal state variables. Additionally, Tekkotsu provides a vision and motion library where higher level perceptions and movements can be specified. The AIBO architecture shown at the right hand side of Figure 1 describes the robot processing modules. To play soccer robots are programmed with a set of behaviors that are activated depending on information read from sensors and state information that includes ball position, game state, localization, number of robots in the field, team strategies, etc.

3.2 Command, Interrogate and Teach Dialogs

In order to demonstrate the human coaching model we have developed and experimented with simple dialogs that let the user: (1) *command* the robot to perform certain actions; (2) *interrogate* the robot specific questions about its state; and (3) *teach* the robot to link a sequence of lower level behaviors into a higher level command such as "*Go get the ball and walk it into the goal*". Videos for these dialogs can be found in [20]. A sample command and interrogate dialog is shown in Table 2.

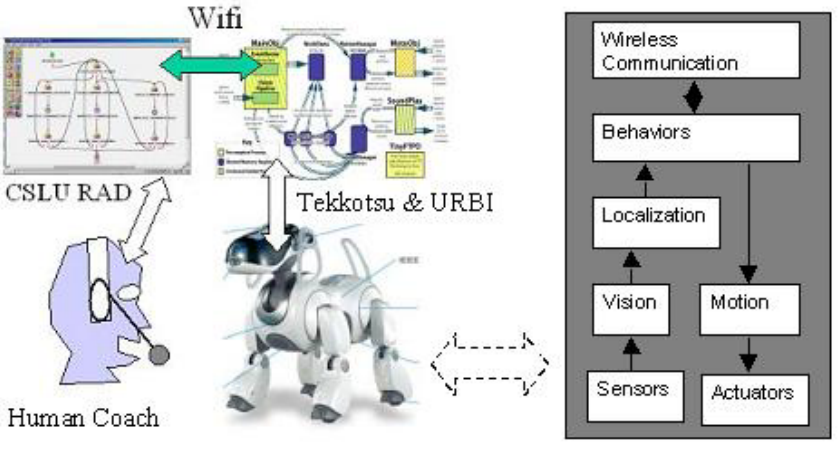

Fig. 1. A schematic view of the CSLU-Tekkotsu/URBI-AIBO system architecture. The diagram to the right shows the internal AIBO processing modules: Sensors, Actuators, Vision, Motion, Localization, Behaviors and Wireless Communication.

Table 2. Example dialog for commanding and interrogating the robot

Robot: *Welcome to the LAFMI Human-Robot Interaction Demonstration. I am the AIBO…*
Human: *Do you see the ball?*
Robot: *Looking for the ball…I don't see anything.*
Human: *Turn right.*
Robot: *Turning right. (Turns right)*
Human: *Do you see the ball?*
Robot: *Looking for the ball… I see the ball.*
Human: *Get the ball.*
Robot: *Going for the ball. (Starts walking to the ball).*
Human: *Stop. Turn right.*
Robot: *Turning right. (Turns right but misses the ball)*
Human: *Back up.*
Robot: *Back up. (Backs up)*
Human: *Turn left.*
Robot: *Turning left. (Turns left but misses the ball)*
Human: *Turn right.*
Robot: *Turning right. (Turns right and catches the ball)*
Human: *Walk.*
Robot: *Walking. (Starts walking, pushing the ball towards the goal)*
Human: *Walk.*
Robot: *Walking. (Keeps walking, pushing the ball and "scoring")*

3.3 Human-Robot Coaching

In pursuing coaching capabilities we utilize the three previous levels of human-robot interaction having been defined in the context of soccer playing robots: command, interrogate and teach. We have defined a set of basic commands, action-only and

action-perception behaviors that can be instructed to the robot. Additionally the robot may be interrogated with state and perception related queries. Finally, these commands form the basis for teaching new behaviors in the soccer playing domain. While different levels of these commands have already been implemented in the AIBO in the context of soccer playing, we are at this point experimenting with them.

Command. We define a set of action-only and action perception commands. Action-only commands i.e. no perception, include: *Stop, Move, Turn*, *Turn Head*, and *Kick Ball.* Depending on the commands, these may include arguments such as magnitude of rotation, and movement in degrees or steps, etc. For example a rotation command would be *Turn 180 degrees* and a movement command would be *Move 4 steps*. It should be noted that at this level commands such as *Kick Ball* would not use any perceptual information, i.e. the resulting kick will be similar (hopefully) to the current robot orientation. We also define a set of action-perception commands requiring the full perception-action cycle, i.e. the action to be performed depends on the current robot perceptions. These commands include: *Kick Ball* with a specified direction; *Reach Ball* moving to a position behind the ball pointing towards the goal; *Initial Position* during game initialization requiring localization in the field; *Pass the Ball* to gently kick the ball to another team robot; *Move to Location* specifying a position in the field where to move; *Search Ball* resulting in robot looking for a ball nearby; *Explore Field* resulting in a more extensive search for the ball; *Defend Goal* resulting in all robots moving close to the goal requiring knowledge of the robot location in the field; *Defend Kick* in trying to block a kick from the other team, requiring knowledge of ball location, and *Attack Goal* similar although opposite in behavior to defending goal.

Interrogation. We define state and perception interrogation commands returning information on current actions or behaviors. State interrogations include for example: *What was your last action,* e.g. kicked the ball; *Why did you take the last action,* e.g., I saw the ball, so I moved towards it; *What is your current behavior*, e.g. I'm searching for the ball; *What is your current role in the game*, e.g. I am the goalie. Perception interrogations include for example: *Do you see the ball* returning e.g. *I do*, *I don't*; *What is your distance to the ball*, returning e.g. *30 centimeters*; *What is your current orientation,* returning e.g. *45 degrees* (in relation to field coordinate system); *What is your current position,* returning e.g. *I am in region 9*; *What is the position of object X* returning an estimate of its position.

Teach. The ultimate goal in human-robot coaching in the context of soccer is being able to positively affect the team performance during a game. While part of this interaction can eventually be carried out by agent coaches inside the robot, it is our goal to define the basic capabilities and communication interactions that human coaches should have. For example, being able to transmit strategy knowledge in the form "*if blocked pass the ball to player behind*". Such a command will modify an internal robot database with "*if possess(ball) and goal(blocked) then pass(ball)*".

4 Conclusions and Discussion

The stated objective of the current research is to develop a generalized approach for human-machine interaction via spoken language that exploits recent developments in

cognitive science - particularly notions of grammatical constructions as form-meaning mappings in language, and notions of shared intentions as distributed plans for interaction and collaboration. In order to do this, we tested human-robot interaction initially with the Event Perceiver system and later on with the Sony AIBOs under soccer related behaviors.

With respect to social cognition, shared intentions represent distributed plans in which two or more collaborators have a common representation of an action plan in which each plays specific roles with specific responsibilities with the aim of achieving some common goal. In the current study, the common goals were well defined in advance (e.g. teaching the robots new relations or new behaviors), and so the shared intentions could be built into the dialog management system.

An initial evaluation period revealed that while technically we had demonstrated command, interrogation and teaching, the user interface ergonomics was somewhat clumsy. In particular the dialog pathways were somewhat constrained, with several levels of hierarchical structure in which the user had to navigate the control structure with several single word commands in order to teach the robot a new relation, and then to demonstrate the knowledge, rather than being able to do these operations in more natural single sentences. In order to address this issue, we reorganized the dialog management where context changes are made in a single step. Also, in order to focus the interactions, we worked around scenarios in which the human and robot collaborate around the shared goal of finding the ball and moving it towards a landmark so that the robot can see both at the same time.

Acknowledgements

Supported by the French-Mexican LAFMI, the ACI TTT Projects in France and the UC-MEXUS CONACYT, CONACYT grant #42440, and “Asociación Mexicana de Cultura” in Mexico.

References

1. Dominey, P.F., Hoen, M., Lelekov, T., Blanc, J.M.: Neurological basis of language in sequential cognition: Evidence from simulation, aphasia and ERP studies. Brain and Language 86(2), 207–225 (2003)
2. Tomasello, M.: Constructing a language: A usage-based theory of language acquisition. Harvard University Press, Cambridge (2003)
3. Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H.: Understanding and sharing intentions: The origins of cultural cognition, Behavioral and Brain Sciences (2006)
4. Dominey, P.F., Boucher, J.D.: Developmental stages of perception and language acquisition in a perceptually grounded robot. Cognitive Systems Research 6(3), 243–259 (2005)
5. Dominey, P.F., Boucher, J.D.: Learning to talk about events from narrated video in a construction grammar framework. Artificial Intelligence 167(1-2), 31–61 (2005)
6. Dominey, P.F., Weitzenfeld, A.: Robot Command, Interrogation and Teaching via Social Interaction. In: IEEE-RAS International Conference on Humanoid Robots, Dec. 6-7, Tsukuba, Japan (2005)

7. Kotovsky, L., Baillargeon, R.: The development of calibration-based reasoning about collision events in young infants. Cognition 67, 311–351 (1998)
8. Siskind, J.M.: Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. Journal of AI Research 15, 31–90 (2001)
9. Steels, L., Baillie, J.C.: Shared Grounding of Event Descriptions by Autonomous Robots. Robotics and Autonomous Systems 43(2-3), 163–173 (2002)
10. Dominey, P.F., Inui, T.: A Developmental Model of Syntax Acquisition in the Construction Grammar Framework with Cross-Linguistic Validation in English and Japanese. In: Proceedings of the CoLing Workshop on Psycho-Computational Models of Language Acquisition, Geneva, pp. 33–40 (2004)
11. Goldberg, A.: Constructions. U Chicago Press, Chicago and London (1995)
12. Bates, E., McNew, S., MacWhinney, B., Devescovi, A., Smith, S.: Functional constraints on sentence processing: A cross linguistic study. Cognition 11, 245–299 (1982)
13. CSLU Speech Tools Rapid application Development (RAD), http://cslu.cse.ogi.edu/toolkit/index.html
14. Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., Osawa, E.: Robocup: The robot world cup initiative. In: Proceedings of the IJCAI-95 Workshop on Entertainment and AI/ALife (1995)
15. Martínez-Gómez, J.A., Medrano, A., Chavez, A., Muciño, B., Weitzenfeld, A.: Eagle Knights AIBO Team, Team Description Paper, VII World Robocup 2005, Osaka, Japan, July 13-17 (2005)
16. Martínez-Gómez, J.A., Weitzenfeld, A.: Real Time Localization in Four Legged RoboCup Soccer. In: Martínez-Gómez, J.A, Weitzenfeld, A. (eds.) Proc. 2nd IEEE-RAS Latin American Robotics Symposium, Sao Luis, Maranhao Brasil, Sept 24-25 (2005)
17. Riley, P., Veloso, M., Kaminka, G.: An empirical study of coaching. In: Distributed Autonomous Robotic Systems 6, Springer, Heidelberg (2002)
18. Kaminka, G., Fidanboylu, M., Veloso, M.: Learning the Sequential Coordinated Behavior of Teams from Observations. In: RoboCup-2002 Symposium, Fukuoka, Japan (June 2002)
19. CMU Tekkotsu: http://www-2.cs.cmu.edu/~tekkotsu/
20. Dominey, P.F., Weitzenfeld, A.: Videos for command, interrogate and teach AIBO robots, ftp://ftp.itam.mx/pub/alfredo/COACHING/
21. Universal Real-time Behavior Interface: http://www.urbiforge.com/