Towards Conversational Artifacts

Toyoaki Nishida[1],

[1] Graduate School of Informatics, Kyoto University, Yoshida-Honmachi Sakyo-ku
606-8501 Kyoto, Japan
nishdia@i.kyoto-u.ac.jp

Abstract. Conversation is a natural and powerful means of communication for people to collaboratively create and share information. People are skillful in expressing meaning by coordinating multiple modalities, interpreting utterances by integrating partial cues, and aligning their behavior to pursuing joint projects in conversation. A big challenge is to build conversational artifacts – such as intelligent virtual agents or conversational robots – that can participate in conversation so as to mediate the knowledge process in a community. In this article, I present an approach to building conversational artifacts. Firstly, I will highlight an immersive WOZ environment called ICIE (Immersive Collaborative Interaction Environment) that is designed to obtain detailed quantitative data about human-artifact interaction. Secondly, I will overview a suite of learning algorithms for enabling our robot to build and revise a competence of communication as a result of observation and experience. Thirdly, I will argue how conversational artifacts might be used to help people work together in multi-cultural knowledge creation environments.

Keywords: Conversational informatics, social intelligence design, information explosion.

1 Prologue

We are in the midst of Information explosion (*Info-plosion*). On the one hand, we often feel overloaded by the overwhelming amount of information, such as too many incoming e-mail messages including spams and unwanted ads. On the other hand, explosively increased information may also lead to a better support of our daily life [1]. Info-plosion has brought about an expectation that dense distribution of information and knowledge in our living space will eventually allow actors to maximally benefit from the given environment being guided by ubiquitous services.

Unfortunately, the latter benefit is not fully there, as one might be often trapped by real world problems, such as being unable to connect the screen of your laptop to the projector. From time to time, the actors might be forced to waste long time to recover from obsolete instructions or lose critical moments due to the lack of timely information provision. Should the knowledge actor fail to complete it in real-time, she or he may not benefit from the knowledge.

A key issue in the information age is knowledge circulation [2]. It is not enough to just deliver knowledge to everybody who needs it. It is critical to keep knowledge

updated, and have it evolve by incorporating ideas and opinions of people. Knowledge need to be circulated among proper people so that they can incorporate contribution from them. Although information and communication technologies provide us with potential keys to success, a wide range of issues need to be addressed, ranging from fundamental problems in communication to cultural sensitivity.

It is quite challenging to address what is called the knowledge grounding problem arising from the fact that information and knowledge on the web are essentially decoupled from the real world, in the sense that they cannot be applied to the real world problems unless the actor properly recognizes the situation and understand how knowledge is associated with it. Propositions decoupled from the real world may cause the "last 10 feet problem", i.e., one might not be able to reach the goal even though s/he is within the 10 feet from there. Computational models need to be built for accounting not only for the process of perceptual knowledge in action but also for the meaning and concept creation in general. We need to address the epistemological aspects of knowledge and build a computational theory of understanding perceptual knowledge we have to live in the real world. How can we do it?

2 Power of Conversation

Conversation plays a critical role in forming grounded knowledge by associating knowledge with real world situations [3]. People are skillful in aligning their behavior to pursuing joint projects in conversation, as Clark characterized conversation as an emergent joint action, to be carried by an ensemble of people [4]. Language use consists of multiple levels, from the signals to joint projects. Various kinds of social interactions are made at multiple levels of granularity. In the middle, speech acts such as requesting for information, proposing solution, or negotiating. In the micro, interaction is coordinated by quick actions such as head gesture, eye gaze, posture and paralinguistic actions. In the macro, long-term social relation building is going, trust-making, social network building, and developing social atmosphere. Occasionally, when they get deeply involved in a discussion, they may synchronize their behavior in an almost unconscious fashion, exhibiting empathy with each other to be convinced that they have established a common understanding.

People are skillful both in expressing meaning by coordinating multiple modalities and in interpreting utterances by integrating partial cues. People not only use signals to control the flow of a conversation, e.g., pass the turn of conversation from one to another but also create or add meaning by making utterances, indicating things in the real world, or demonstrating aspects of objects under discussion. Kendon regarded gestures as a part of speaker's utterances and conducted a descriptive analysis of gesture use by investigating in detail how speech and gesture function in relation to one another [5]. McNeill discussed the mental process for integrated production of gesture and words [6].

3 Conversational Artifacts

Conversational artifacts are autonomous software or hardware capable of talking with people by integrating verbal and nonverbal means of communication. The role of conversational artifacts is to mediate the flow of conversational content among people.

There is a long history of development for embodied conversational agents or intelligent virtual agents [7], [8]. Our group has been working on embodied conversational agents and conversational robots [9-14].

As the more sophisticated agents are being built, the methodology has shifted from the script/programming-based to data-driven approaches, for we need to gain more detailed understanding of communicative proficiency people show in conversation. The data-driven approach consists of two stages: the first stage for building a conversation corpus by gathering data about inter-human conversation and the second stage for generating the behavior of conversational artifacts from the corpus. WOZ (Wizard-of-Oz) is effective in collecting data in which a tele-operated synthetic character or robot are used to interact with experiment participants.

In order for this approach to be effective, two technical problems need to be solved. The first is to realize the "human-in-the-artifacts" feeling. In WOZ experiments, we employ experiment participants to operate conversational to collect how the conversational artifacts should act in various situations in conversation. In order for these WOZ experiments to be useful, the experiment participants should feel and behave as if she were the conversational artifact. Thus, the WOZ experiment environment should be able to provide experiment participants with the situational information the conversational artifact obtains and operate the conversational artifact without difficulty. The second is to develop a method of effectively producing the behaviors of the conversational artifact from the data collected in the WOZ experiments. I will address these issues in the following two sections.

4 Immersive WOZ Environment with ICIE

Our immersive WOZ environment provides the human operator with a feeling as if s/he stayed "inside" a conversational artifact to receive incoming visual and auditory signals and to create conversational behaviors in a natural fashion [15]. At the human-robot interaction site, a 360-degree camera is placed near the robot's head, which can acquire the image of all directions around it. The image captured by the 360-degree camera is sent to the operator's cabin using TCP/IP. The WOZ operator's cabin is in the cylindrical display, which is a set of large-sized displays which are circularly aligned. The current display system uses eight 64-inch display panels arranged in a circle with about 2.5 meters diameter. Eight surround speakers are used to reproduce the acoustic environment. The WOZ operator stands in the cylindrical display and controls the robot from there. The image around the robot is projected on an immersive cylindrical display around the WOZ operator. This setting gives the operator exactly the same view as the robot sees. When a scene is displayed on the full screen, it will provide a sense of immersion.

The WOZ operator's behavior, in turn, is captured by a range sensor to reproduce a mirrored behavior of the robot. We realize accurate and real-time capturing of the operator's motion by using a range sensor and enable the operator to intuitively control the robot according to the result of the capturing. We make the robot take the same poses as the operator does by calculating the angles of the operator's joints at every frame. We can control NAO's head, shoulders, elbows, wrists, fingers, hip joints, knees, and ankles, and we think they are enough to represent basic actions in communication. The sound on each side of the WOZ operator is gathered by microphones and communicated via network so that everyone can hear the sound of the other side.

5 Learning by Mimicking

Learning by mimicking is a computational framework for producing the interactive behaviors of conversational artifacts from a collection of data obtained from the WOZ experiments. In the framework of learning by mimicking, a human operator is guiding a robot (actor) to follow a predefined path in the ground using free hand gestures. Another learner robot watches the interaction using sensors attached to the operator and the actor and learns the action space of the actor, the command space of the operator and the associations between commands (gestures) and actions. This metaphor characterizes our approach to developing a fully autonomous learner, which might be contrasted with another approach to manually producing the behavior of conversational artifacts probably partially using data mining and machine learning techniques. Currently, we concentrate on nonverbal interactions though we have started on integrating verbal and nonverbal behaviors. We have developed a suite of unsupervised learning algorithms for this framework [16][17].

The learning algorithm can be divided into four stages:

1) the discovery stage on which the robot discovers the action and command space;
2) the association stage on which the robot associates discovered actions and commands generating a probabilistic model that can be used either for behavior understanding or generation;
3) the controller generation stage on which the behavioral model is converted into an actual controller to allow the robot to act in similar situations; and
4) the accumulation stage on which the robot combines the gestures and actions it learned from multiple interactions.

6 Application to Multi-Cultural Knowledge Creation

Cultural factors might come into play in globalization. Based on the work on cross-cultural communication [18], we are currently investigating how difficulties in living in a different culture are caused by different patterns of thinking, feeling and potential actions. We are building a *simulated crowd*, a novel tool for allowing people to practice culture-specific nonverbal communication behaviors [19].

We have started a “cross-campus exploration” project aiming at prototyping a system that allows the user (e.g., in the Netherlands) to explore (probably in a RPG fashion) a virtualized university campus possibly in a different culture (e.g., in Japan), or use a tele-presence robot to meet people out there. It will permit the user to experience with interacting with people in a different culture or even actually. Technologies for conversational artifacts will play a significant role in these applications.

References

1. Kitsuregawa, M., Nishida, T.: Special Issue on Information Explosion. New Generation Computing. 28(3), 207--215 (2010)
2. Nishida, T.: Social Intelligence Design for Cultivating Shared Situated Intelligence. In: GrC 2010: 369-374 (2010)
3. Nishida, T. (ed.): Conversational Informatics: an Engineering Approach, John Wiley & Sons Ltd, London (2007)
4. Clark, H.H.: Using Language, Cambridge University Press (1996)
5. Kendon, A.: Gesture, Cambridge University Press (2004)
6. McNeill, D.: Gesture and Thought, The University of Chicago Press (2005)
7. Cassell, J., Sullivan, J., Prevost, J., and Churchill, E. (eds.): Embodied Conversational Agents, The MIT Press (2000)
8. Prendinger, H. and Ishizuka, M. (eds.): Life-like Characters -- Tools, Affective Functions and Applications, Springer-Verlag (2004)
9. Kubota, H., Nishida, T., and Koda, T.: Exchanging Tacit Community Knowledge by Talking-virtualized-egos, In: Proceedings of Agent 2000, 285--292 (2000)
10. Nishida, T.: Social Intelligence Design for Web Intelligence, Special Issue on Web Intelligence, IEEE Computer 35(11), 37--41 (2002)
11. Okamoto M., Nakano, Y.I., Okamoto, K., Matsumura, K., and Nishida, T.: Producing Effective Shot Transitions in CG Contents based on a Cognitive model of User Involvement, IEICE Transactions of Information and Systems Special Issue of Life-like Agent and Its Communication, IEICE Trans. Inf. & Syst. E88-D(11), 2623--2532 (2005)
12. Huang, H.H., Cerekovic, A., Pandzic, I., Nakano, Y, and Nishida, T.: The Design of a Generic Framework for Integrating ECA Components, In: Proceedings of 7th International Conference of Autonomous Agents and Multiagent Systems (AAMAS08), Estoril, Portugal, 128—135 (2008)
13. Huang, H.H., Furukawa, T., Ohashi, H., Nishida, T., Cerekovic, A., Pandzic, I.S., Nakano, Y.I.: How Multiple Concurrent Users React to a Quiz Agent Attentive to the Dynamics of their Game Participation. In: AAMAS 2010: 1281--1288 (2010)
14. Nishida, T., Terada, K., Tajima, T., Hatakeyama, M., Ogasawara, Y., Sumi, Y., Yong, X., Mohammad, Y.F.O., Tarasenko, K., Ohya, T., and Hiramatsu, T.: Towards Robots as an Embodied Knowledge Medium, Invited Paper, Special Section on Human Communication II, IEICE TRANSACTIONS on Information and Systems E89-D(6), 1768--1780 (2006)
15. Ohashi, H., Okada, S., Ohmoto, Y., and Nishida, T.: A Proposal of Novel WOZ Environment for Realizing Essence of Communication in Social Robots, Presented at: Social Intelligence Design 2010 (2010)
16. Mohammad, Y.F.O., Nishida, T., and Okada, T.: Unsupervised Simultaneous Learning of Gestures, Actions and their Associations for Human-Robot Interaction. In: IROS 2009: 2537--2544 (2009)

17. Mohammad, Y.F.O. and Nishida, T.: Learning Interaction Protocols using Augmented Baysian Networks Applied to Guided Navigation, Taipei, Taiwan, Presented at: IROS 2010 (2010)
18. Rehm, M., Nakano, Y.I., André, E., and Nishida, T.: Culture-Specific First Meeting Encounters between Virtual Agents. In: IVA 2008: 223--236 (2008)
19. Thovuttikul, S., Lala, D., Ohashi, H., Okada, S., Ohmoto, Y., and Nishida, T.: Simulated Crowd: Towards a Synthetic Culture for Engaging a Learner in Culture-dependent Nonverbal Interaction, Presented at: 2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction, Stanford University, USA (2011)