Human–machine metacommunication towards development of a human-like agent: A short review

Akinori Ito*

Graduate School of Engineering, Tohoku University,
6–6–5 Aramaki aza Aoba, Aoba-ku, Sendai, 980–8579 Japan

Abstract: This article briefly reviews the research works related to metacommunication. Metacommunication is a term meaning “communication on communication,” which is related to marginal communication such as conveying recognition, comprehension, and evaluation of an interlocutor’s words. Herein, several research works are reviewed from the metacommunication point of view.

Keywords: Spoken dialog systems, Human-like agent, Metacommunication

PACS number: 43.72.Kb [doi:10.1250/ast.41.166]

1. INTRODUCTION

Artificial intelligence (AI) technology has drastically advanced in this decade. The development of an AI agent that behaves like a human is one of the ultimate goals of AI research. There has been much effort devoted to developing such an agent. However, it is still difficult to develop a completely human-like agent. Thus, researchers have developed technologies for increasing human likeness from various aspects, such as learning ability [1,2], communication and reasoning in a game [3], emotional expression in a chatbot [4], and appearance [5]. This article addresses the communication channel between a machine and a user.

Most spoken dialog systems assume that the user of the system is ready to talk with the system, and the system gives a prompt utterance to the user, expecting the user to reply to the prompt immediately. This assumption makes sense for task-oriented systems such as interactive voice response systems [6] or an information kiosk [7]. However, current spoken dialog systems, such as “smart speakers” and “intelligent personal assistants” [8], are not only task-oriented voice interfaces but also “chatbots” [9]. When used as a non–task-oriented chatbot, the user does not necessarily have a clear goal of the conversation. To realize a spoken dialog system that can make “human-like” conversation under such a situation, we must make the system be conscious of the relationship between the system and the user. In other words, a “communication channel” between the system and the user is necessary.

When we make conversation with other persons, we are always conscious of the channel to an interlocutor, and we interchange information on the “state” of the channel, such as the floor and turn-taking, what topic is preferred, and whether the listener is actually listening to the speaker’s words or not. These kinds of interchanges are called “metacommunication.” This article reviews the research work on the human–machine interaction from the metacommunication point of view, and the future direction of human–machine metacommunication research towards the realization of a human-like conversational agent is discussed.

2. CONCEPT OF METACOMMUNICATION

2.1. What is Metacommunication?

In this article, the term “metacommunication” denotes all aspects of communication related to the communication itself.

The word “metacommunication” was coined by Bateson [10]. He defined the word metacommunication as “communication that refers to communication.” Since there is no consensus on the definition of metacommunication, some researchers refer to only verbal communication on communication (such as confirmation utterances) as “metacommunication” and nonverbal communication as “paracommunication” [11], while others refer to all communicative acts that refer to communication as “metacommunication” [12]. Herein the latter stance is adopted. “Metacommunication” means not only the intentional communication but also the unintentional expression of one’s state and the perception of the relationship from the expression, such as the physical distance between two
persons, gaze and facial expressions, filler words, and disfluency.

2.2. Why Metacommunication is Important

Consider a situation where a user is verbally ordering something to a robot, but the robot does not respond to the voice command. If the robot does not take any action, it is difficult for the user to understand what is happening. There are many possibilities of the cause, such as the robot has failed to recognize the voice command because of environmental noise, the robot cannot understand the command because the user used out-of-vocabulary words, or the microphone of the robot is malfunctioning.

If the robot presents behavior such as saying “I could not hear you” or emits a beep sound indicating a recognition failure, the user can take an appropriate next action, such as repeating the voice command. Such action of the robot is unrelated to the target task of the robot. Instead, it is a the robot behavior to express the status of the communication channel between the user and the robot. This is an example of human–machine metacommunication.

Not only the “machine-to-human” case, but the “human-to-machine” case, where a machine understands the human’s cognition of the communication channel, is also useful. For example, when a person moves toward a communication robot, it might mean that the person wants to talk with the robot. Considering a communication robot that chats with a person, it is not an easy task to determine when the robot should start and finish the talk. The system needs to observe the user’s multimodal behavior and perceive cues of metacommunication to judge whether the user wants to continue the talk or not.

3. EXAMPLES OF METACOMMUNICATION

There are many aspects of metacommunication. The following issues are especially important in realizing human-like spoken dialog systems.

3.1. Appearance

The appearance of a system, such as its height and figure (such as human, robot, animal, or cartoon), affects the social relationship between the system and the user. Thus, the system’s appearance should be designed in accordance with the purpose of the system. Martini et al. [5] investigated the relationship between the appearance of a robot and the impression of humans regarding the robot. From the point of view of the spoken dialog system, Bickmore and Cassell [13] described the effect of displaying the figure of the agent of the spoken dialog system. Their result was that the figure was effective when the dialog was simple, whereas it had no effect when the user conducted a complicated dialog. Miyake and Ito [14] compared the linguistic feature of the speech in dialogs with and without the agent’s figure.

In addition to the figure of the agent, the height of the agent also affects the psychological relationship between the agent and a human. As an effect of a person’s height, it is known that taller persons have higher income [15]. Furthermore, in video-based communication, placing the display at a higher position affects the conversation [16]. These results suggest that the physical height of an agent is an important factor of a spoken dialog system. Hiroi and Ito investigated the best height of a conversational robot and concluded that the best height was about 30 cm lower than the eye level of humans [17].

3.2. Establishment and Termination of the Conversation

Starting and finishing the conversation are important issues for spoken dialog systems, but most systems avoid this issue by assuming that the user has the initiative in establishing and finishing the dialog. When a user greets the system, it means that the user wants to make new conversation with the system. In the situation where the system establishes the dialog, however, it is important for the system to know when it can talk to the user. Hudson et al. proposed to estimate the “interruptibility” (to what degree the system can interrupt the user to start a dialog) of a user on the basis of sensor information [18]. It is also important for a robot to know how to approach the user and initiate conversation [19].

3.3. Keeping a Conversation Channel Connected

While the system and the user are talking, it is required for both sides of the conversation to keep the conversation channel connected. If one of the talkers were to suddenly become silent, the conversation channel would be closed and the conversation would finish. Proper turn-taking and backchannel help to keep the conversational channel connected [20]. Backchannel prediction [21] is an important issue for realizing natural human–machine conversation. It is also known that synchrony of the gaze [22] and body movement [23] is observed during a conversation. These multimodal behaviors should be considered when developing a visual conversational agent or a conversational robot.

3.4. Noise-Adaptive Behavior Control

Environmental noise is one cause of hindering smooth interaction. Usually, noise-robust speech recognition methods are employed to cope with environmental noise [24]. As an alternative way, several systems change dialog acts adaptively in accordance with the noise level. The dialog system developed by Ito et al. [25] changes the vocabulary and dialog strategy so that the system uses a smaller
vocabulary to increase the recognition rate. Chandramohan and Pietquin incorporated the noise level into the features of the Markov decision process to determine the next dialog act [26]. These kinds of behavior control can also be seen as examples of metacommunication.

3.5. Awareness and Understanding of the Conversation

In a conversation, a speaker observes the interlocutor’s behavior and perceives whether the interlocutor understands the speaker’s words, or whether he/she can answer the speaker’s question or not. Confirmation is the most basic metacommunication employed by most task-oriented dialog systems. Confirmation is a kind of metacommunication because a confirmation is not an activity to convey new information; rather, it is an activity to share already-conveyed information. Most task-oriented spoken dialog systems perform confirmation after collecting all information needed to accomplish the task. Some systems employ a dialog strategy that adaptively determines whether the systems should make the confirmation or not. For example, Ultes et al. [27] developed a dialog system that observed the “interaction quality (IQ),” which indicates how smooth the dialog is and determines the confirmation behavior in accordance with the value of IQ.

The ability of a human to know whether the interlocutor knows specific information is called “feeling of another’s knowing” (FOAK) [28]. This kind of ability to estimate the metacognitive state of the interlocutor is indispensable for a dialog system to realize a human-like conversational agent. There are only a few research works in this field, for example, Chiba et al. attempted to discriminate what kind of difficulty the speaker has when the speaker does not give a response [29]. Recognizing the awareness and understanding of a speaker is especially important in multiparty conversation [30].

3.6. Interest and Evaluation of Conversation

A human speaker is also able to perceive whether the interlocutor is interested in the conversation [31], and whether he/she agrees with what the speaker says. Not only understanding the user’s behavior but also expressing the system’s mental status to the user is an important issue [32]. Buschmeier and Kopp [33] developed a system for estimating the listener’s mental state using a Bayesian network. The states estimated by this system include the understanding, acceptance, or agreement with the speaker’s statement.

Emotion in conversation is a similar concept to the interest and evaluation of the conversation. Emotion is said to be useful in enhancing the rapport between the user and the system [34]. Yang et al. investigated the user impression of dialog systems that employ animated agents with or without emotional expressions [35]. They found that the agent showing emotional behavior was perceived to be more intelligent, more empathetic and more conscious.

4. TOWARDS A HUMAN-LIKE CONVERSATIONAL AGENT

There have been many research attempts to develop human-like conversational agents [36], such as a humanoid robot [37] and animated face [38]. These works focus on how to replicate human behavior precisely. In addition to the precise replication of human behavior, it seems to be important to consider the meaning of the behaviors of an agent from a metacommunication point of view. It is known that not only a human-like appearance but also natural behavior such as eye contact affects the perception of the agent by humans [39]. Changing the behavior of a robot (the “embodied cues”) while interacting with a human affects the performance of the human [40]. These research works reveal the importance of the metacommunicative behavior of an agent.

REFERENCES

[1] N. Li, N. Matsuda, W. W. Cohen and K. R. Koedinger, “Integrating representation learning and skill learning in a human-like intelligent agent,” Artif. Intell., 219, 67–91 (2015).
[2] B. M. Lake, T. D. Ullman, J. B. Tenenbaum and S. J. Gershman, “Building machines that learn and think like people,” Behav. Brain Sci., 40, E253 (2017).
[3] N. Nakamura, M. Inaba, K. Takahashi, F. Toriumi, H. Osawa, D. Katagami and K. Shinoda, “Constructing a human-like agent for the Werewolf Game using a psychological model based multiple perspectives,” Proc. IEEE Symp. Computational Intelligence (SSCI), pp. 1–8 (2016).
[4] R. E. Banchs, “On the construction of more human-like chatbots: Affect and emotion analysis of movie dialogue data,” Proc. APSIPA ASC, pp. 1364–1367 (2017).
[5] M. C. Martini, C. A. Gonzalez and E. Wiese, “Seeing minds in others — Can agents with robotic appearance have human-like preferences?” PloS ONE, 11(1), e0146310 (2016).
[6] R. Corkey and L. Parkinson, “Interactive voice response: Review of studies 1989–2000,” Behav. Res. Methods Instrum. Comput., 34, 342–353 (2002).
[7] S. Kopp, L. Gesellensetter, N. C. Krämer and I. Wachsmuth, “A conversational agent as museum guide—Design and evaluation of a real-world application,” in Intelligent Virtual Agents (Springer, Berlin, Heidelberg, 2005), pp. 329–343.
[8] G. López, L. Quesada and L. A. Guerrero, “Alexa vs. Siri vs. Cortana vs. Google Assistant: A comparison of speech-based natural user interfaces,” Proc. Int. Conf. Applied Human Factors and Ergonomics, pp. 241–250, Springer (2017).
[9] R. Dale, “The return of the chatbots,” Nat. Lang. Eng., 22, 811–817 (2016).
[10] R. W. Mitchell, “Bateson’s concept of metacommunication” in play,” New Ideas Psychol., 9, 73–87 (1991).
[11] B. D. Beitman and G. I. Viamontes, “Unconscious role-induction: Implications for psychotherapy,” Psychiatr. Ann., 37, 259–265 (2007).
[12] H. Bahri and R. Williams, “Interconnectivity and metacommunication,” Train. Lang. Cult., 1(3), pp. 8–30 (2017).
A. ITO: HUMAN–MACHINE METACOMMUNICATION

[13] T. Bickmore and J. Cassell, “Social dialogue with embodied conversational agents,” in Advances in Natural Multimodal Dialogue Systems, J. C. J. van Kuppevelt, L. Dybkjær and N. O. Bernsen, Eds., volume 30 of Text, Speech and Language Technology (Springer, Dordrecht, 2005), Chap. 1, pp. 23–54.

[14] S. Miyake and A. Ito, “A spoken dialogue system using virtual conversational agent with augmented reality,” Proc. APSIPA ASC, pp. 1–4 (2012).

[15] T. Judge and M. D. Cable, “The effect of physical height on workplace success and income: Preliminary test of a theoretical model,” J. Appl. Psychol., 89, 428–441 (2004).

[16] W. Huang, S. J. Olson and M. G. Olson, “Camera angle affects dominance in video-mediated communication,” Proc. Conf. Human Factors in Computing Systems (CHI’02), pp. 716–717 (2002).

[17] Y. Hiroi and A. Ito, “Effect of the size factor on psychological threat of a mobile robot moving toward human,” KANSEI Eng. Int., 8(8), pp. 51–58 (2009).

[18] S. Hudson, J. Fogarty, C. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. Lee and J. Yang, “Predicting human interruptibility with sensors: A Wizard of Oz feasibility study,” Proc. SIGCHI Conf. Human Factors in Computing Systems, pp. 257–264 (2003).

[19] S. Satake, T. Kanda, D. F. Glas, M. Imai, H. Ishiguro and N. Hagita, “How to approach humans?—Strategies for social robots to initiate interaction,” Proc. 4th ACM/IEEE Int. Conf. Human-Robot Interaction (2009).

[20] R. Poppe, K. P. Truong and D. Heylen, “Perceptual evaluation of backchannel strategies for artificial listeners,” Auton. Agents Multi-Agent Syst., 27, 235–253 (2013).

[21] M. Mueller, D. Leuschner, L. Briem, M. Schmidt, K. Kilgour, S. Stueker and A. Waibel, “Using neural networks for data-driven backchannel prediction: A survey on input features and training techniques,” Proc. Int. Conf. Human-Computer Interaction, pp. 329–340, Springer (2015).

[22] F. Cummins, “Gaze and blinking in dyadic conversation: A study in coordinated behaviour among individuals,” Lang. Cogn. Process., 27, 1525–1549 (2012).

[23] P. Paxton and R. Dale, “Frame-differencing methods for measuring bodily synchrony in conversation,” Behav. Res. Methods, 45, 329–343 (2013).

[24] J. Li, L. Deng, Y. Gong and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” IEEE/ACM Trans. Audio Speech Lang. Process., 22, 745–777 (2014).

[25] A. Ito, T. Oba, T. Konashi, M. Suzuki and S. Makino, “Selection of optimum vocabulary and dialog strategy for noise-robust spoken dialog systems,” IEICE Trans. Inf. Syst., E91-D, 538–548 (2008).

[26] S. Chandramohan and O. Pietquin, “User and noise adaptive dialogue management using hybrid system actions,” Proc. Int. Workshop Spoken Dialogue Systems Technology, pp. 13–24, Springer (2010).

[27] S. Ultes, H. Dikme and W. Minker, “Dialogue management for user-centered adaptive dialogue,” in Situated Dialog in Speech-based Human–Computer Interaction (Springer, Cham, 2016), pp. 51–61.

[28] S. E. Brennan and M. Williams, “The feeling of another’s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers,” J. Mem. Lang., 34, 383–398 (1995).

[29] Y. Chiba, T. Nose and A. Ito, “Cluster-based approach to discriminate the user’s state whether a user is embarrassed or thinking to an answer to a prompt,” J. Multimodal User Interfaces, 11, 185–196 (2017).

[30] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro and N. Hagita, “Footing in human-robot conversations: How robots might shape participant roles using gaze cues,” Proc. 4th ACM/IEEE Int. Conf. Human Robot Interaction, pp. 61–68, ACM (2009).

[31] Y. Chiba, T. Nose and A. Ito, “Analysis of efficient multimodal features for estimating user’s willingness to talk: Comparison of human–machine and human-human dialog,” Proc. APSIPA ASC, pp. 428–431 (2017).

[32] D. Bohus and E. Horvitz, “Managing human-robot engagement with forecasts and... um.. hesitations,” Proc. 16th Int. Conf. Multimodal Interaction, ICMI ’14, pp. 2–9, ACM, New York (2014).

[33] H. Buschmeier and S. Kopp, “Using a Bayesian model of the listener to unveil the dialogue information state,” Proc. 16th Workshop Semantics and Pragmatics of Dialogue (SemDial 2012), (2012).

[34] J. C. Acosta and N. G. Ward, “Achieving rapport with turn-by-turn, user-responsive emotional coloring,” Speech Commun., 53, 1137–1148 (2011).

[35] Y. Yang, X. Ma and P. Fung, “Perceived emotional intelligence in virtual agents,” Proc. 2017 CHI Conf. Extended Abstr. Human Factors in Computing Systems, pp. 2255–2262, ACM (2017).

[36] L. C. Klopfenstein, S. Delpiori, S. Malatini and A. Bogliolo, “The rise of bots: A survey of conversational interfaces, patterns, and paradigms,” Proc. 2017 Conf. Designing Interactive Systems, pp. 555–565, ACM (2017).

[37] B. Mutlu, J. Forlizzi and J. Hodgins, “A storytelling robot: Modeling and evaluation of human-like gaze behavior,” Proc. Int. Conf. Humanoid Robots, pp. 518–523 (2006).

[38] S. Mariotoyand and C. Busso, “Generating human-like behaviors using joint, speech-driven models for conversational agents,” IEEE Trans. Audio Speech Lang. Process., 20, 2329–2340 (2012).

[39] A. Abubshait and E. Wiese, “You look human, but act like a machine: Agent appearance and behavior modulate different aspects of human–robot interaction,” Front. Psychol., 8, 1393 (2017).

[40] B. Mutlu, “Designing embodied cues for dialogue with robots,” AI Mag., 32(4), pp. 17–30 (2011).