Cascaded Dialogue Modelling for Situated Human-Robot Interactions

Kristiina Jokinen
AI Research Center, AIST Tokyo Waterfront, Japan
kristiina.jokinen@aist.go.jp

Abstract

The paper discusses new requirements for dialogue modelling that situated human-robot interactions and increasingly demanding smart contexts impose on dialogue systems. Based on Constructive Dialogue Modelling, the paper describes human communication enablements (contact, perception, understanding and reaction) and their modelling as a hierarchy which supports the agent to produce appropriate dialogue behaviour. The hierarchy is compared to the subsumption architecture for mobile robots, and a cascaded model with multiple parallel yet hierarchical dialogue processing tasks is proposed as a model for context-aware dialogue management.

1 Introduction

Interactive robot agents are being rapidly developed due to advances in social robotics, and likely to enter into use in great numbers in the domains such as healthcare, home services, nursing, caregiving, education, etc. [1]. Speaking heads and mobile chatbots are already in everyday usage and part of many similar services [2], while robot agents are expected to handle tasks that also include physical action besides verbal interaction. Moreover, in the area of industrial robotics, the robot’s ability to assist humans verbally and explain actions in a cooperative manner is seen as an important extension to robot functionality [3]. Consequently, dialogue interactions with robot agents are likely to increase in frequency and complexity, as the robot should not only provide useful information, but instruct and advise human users, explain its behaviour, and converse about interesting topics in natural language. In fact, the vision of Society 5.0 [4] sees AI agents embedded seamlessly into society. They are connected and capable of communicating with each other and with humans, and by enabling interaction between AI and the human users, they bring a human-centred dimension to AI solutions in service design and decision making.

Interactive robots are social by default. Spoken interaction creates expectations for the robot’s appropriate social behaviour, which includes not only its ability to provide relevant dialogue responses but also ability to address the user’s emotions and affective state [11]. Dialogue management presupposes models for turn-taking, topic tracking, feedback, empathy, etc., while safe and ethical data management concerns privacy issues which especially in caretaking and healthcare applications are crucial issues: the robot agent should deal with sensitive information in a way that is regarded as tactful and trustworthy [17]. In an earlier paper [14] we suggested that the development of human robot interaction produces a new category of interactive agents in the world: that of boundary crossing robots. Such robot agents facilitate the crossing of boundaries in how we conceptually categorize talking robots as agents, and the crossing of boundaries in how we see robots at work environments as friendly co-workers and members of a team. Context-aware social robots which can align themselves with the users’ needs and intentions and support trustworthiness, friendliness and usability in the interactions, may function as such boundary crossing agents and positively influence the users’ view of artificial systems over time [13][19][22].

Social robots can be defined from two perspectives: on one hand, they are sophisticated computer tools, and on the other hand, they are communicating agents. As a computer device, the social robot can offer a multimodal interface to digital data, and its advanced technical properties allow quick processing of large amounts of data and a possibility to connect to various sensors, devices, and smart environments. As an agent whose defining functionality is its dialogue capability, the social robot is expected to conduct natural language interaction with the user in a multimodal manner, and the human-like appearance of a humanoid robot also reinforces such expectations [11][15][18]. A social robot’s capabilities can be referred to by the notion of 4E Cognition which in cognitive robotics refers to cognition that is embodied, embedded, enactive and extended [21]. Although discussion of the assumed cognitive capabilities of robot agents is beyond the scope of this paper, 4E Cognition succinctly describes the novel situation in which humans find themselves, and which affects human interaction as well as the roles which social robots can take in interaction management, therefore requiring restructuring and recategorizing of our views of the world.

The paper focuses on the increasing demands on dialogue architectures in situated human-robot interactions. Section 2 describes Constructive Dialogue Model and its relevance to extended digital dialogue contexts. Section 3 discusses subsumption architectures for CDM and a cascaded model of dialogue processing where tasks are separated yet interleaved as enablements on different levels. Section 4 discusses further future challenges for social and situated robots, and Section 5 concludes the paper.
2 Constructive Dialogue Model (CDM)

2.1 Extended dialogue contexts

In smart environments with interconnected devices, dialogue management between humans and robot agents is becoming more complex, as the context is extended from tangible micro-environments to intangible macro-environments [11]. Interactions are no longer always face-to-face, and the partners need not share their immediate context nor be able to touch each other nor follow each other’s focus of attention.

However, for smooth interactions both in micro- and macro-contexts it is important that the partners coordinate their actions and ensure mutual understanding of the relevant aspects that are crucial for achieving their goals. Dialogue context serves as an immediate memory, and through their interaction, the agents create a shared dialogue context to ground observed dialogue acts and topics. Grounding refers to the act of confirming information with the partner to construct mutual understanding of the goals and intentions [8][9]. In dialogue modelling, grounding is usually done by means of asking and giving feedback on the presented information, producing topically relevant contributions, and by multimodal signalling. In robotics, the term refers to the linking of actions and objects in the physical environment. In both cases, however, grounding refers to sharing and confirming of the presented information so as to construct a joint context.

In macro-environments, the context, which is not immediately sharable needs to be considered in the process of grounding. The agent carries a significant burden of information that is not directly available to the human partner and may not even be shared with the partner (for example due to privacy issues), and thus the dialogue manager needs functionality that enables it to explain its knowledge to the user as well as giving metalevel explanations of any restrictions on data sharing. Network connections can further constraint interactions by delaying the partner’s reactions, and the robot’s dialogue management needs to take such delays into account. Consequently, in digital and virtual environments grounding needs to include network-related contextual information, and a social robot must be equipped with a dialogue manager which takes context-dependent contact into account.

Moreover, the context of the interactions is not static but changes dynamically. Face-to-face interactions have continuity in the creation of common context since the shared physical space allows the participants to be in the same observable environment. However, this is not the case in macro-environments where the context is available through a voice or video channel only (connected virtual environments). The partner’s immediate context is not directly accessible, and it cannot be modified or intervened in by the other agent’s physical actions.

Finally, network data is to be distinguished from the data received through the sensors, such as audio-visual or environmental devices connected to the robot agent to enhance its perception capabilities. While network functioning affects the possibility of the participants to be in contact and create joint context, sensor information is linked to the ability to obtain information from the physical environment, and thus extends the agent’s perception of the environment in which the interaction takes place.

These signals operate on a different level than just being in contact: sensor data needs to be interpreted and correlated with meaningful domain concepts and with the tasks that the robot agent is involved in, i.e. the perceptions need to be grounded. They impose challenges on dialogue management, as the process requires fusion of signal streams and multimodal processing to correlate and integrate sensor data with visual and vocal information, and to create a consistent understanding of the physical and dialogue context.

Interactions in smart-home environments are thus novel situations requiring rich a capability of multimodal information management as well as a versatile repertoire of dialogue models from the robot agent. However, dialogue partners are no longer only humans with natural language conversational skills but can also include humanoid social robots as well as smart objects that communicate through environmental properties (temperature, humidity, etc.). Requirements for dialogue modelling can thus vary from the basic goals of being in contact and perceiving the partner’s intention to communicate to full conversational communication. Analogously, human involvement can range from being fully in charge of the task (e.g. monitoring a situation) to a less commanding role with the robot agent carrying out more demanding tasks.

2.2 Constructing dialogue interaction

The Constructive Dialogue Model (CDM) is a framework for modelling the construction and co-creation of dialogues in natural conversational interactions with dialogue partners [10]. Following the ideas of Communicative Activity Analysis [5], CDM defines four enablements of communication: Contact (C), Perception (P), Understanding (U) and Reaction (R), see Fig. 1.

For Figure 1: Enablements for communication, modified from [11].

The CPUR quadruple forms a set of parallel yet hierarchically organised enablements layers, where the lower level enablements are prerequisites for the higher-level ones to operate successfully. For instance, if there is no contact and perception, it is not possible to communicate with the partner, and if there is contact and perception but not understanding, interaction will fail because there is no relevant reaction. Successful communication requires continuous monitoring that the CPUR enablements are valid, and communication can proceed as intended.

The cycle of interaction forms the basis for grounding, which, as already mentioned, is to construct joint context in which the current dialogue is interpreted. The dialogue context is usually partitioned into three parts concerning the agent’s own knowledge, the agent’s beliefs about the partner’s knowledge, and the agent’s beliefs about the
mutual knowledge. The dialogue thus progresses in cooperation and is co-created by the participants’ contributions through the CPUR cycle. It is also stored in the dialogue history (participant’s long-term memory) as an experience of interaction with the particular partner, and thus each dialogue interaction will also have influence on the future interactions with the same partner.

2.3 CDM Architecture

The CDM enablements can be implemented in a dialogue manager model as shown in Fig. 2. The overall goal of the dialogue manager is to provide appropriate responses to the user’s utterances through the analysis of the new information conveyed by the user’s utterance, and reasoning about its meaning in the context with respect to the agent’s own goals. Information is represented by topics and dialogue acts, the former roughly catching the content of the utterance and the later the speaker’s intention. Understanding is performed by various submodules which take care of specific subgoals related to natural language understanding, intent classification, dialogue policy, planning, etc. The Understanding modules aim to interpret the recognized signals as a meaningful message in the given context. This requires that the participants are mutually aware of each other and of their intention to communicate, and their closeness is checked by the Contact modules which detect and receive communicative signals. The processing of the signals as meaningful communicative intents is done by Perception modules that recognize the signals (auditory, visual, tactile, environmental, etc.) as containing meaningful units. As mentioned, an important enablement of the interaction is the response to the new state of the context as interpreted by the CPU levels. The Reaction modules execute the agent’s reaction which appears as a response to the partner who will then interpret it according to the CPU levels and produce a new reaction, which the agent will process through the CPUR cycle. It is good to notice that planning and language generation modules are regarded as part of the Understanding, not Reaction. This is due to the view that understanding is already response generation since it is the basis for linking the partner’s utterance to the dialogue context, or, to put it the other way round, the answer to the question “where does response planning start?” is that it starts already in the understanding of what the partner said and the response is simply feedback to the partner of one’s understanding of the new information [8].

CDM as such is not commit to rule-based or neural dialogue management. Modularity of dialogue architectures has long been a major issue in dialogue management. As the tasks and interactions for a robot agent become more complex, single end-to-end models are too coarse and too difficult to maintain good control of the various aspects of the dialogue flow in order to support natural, user-adapted and long-term interactions. A hybrid architecture seems a better alternative. Moreover, since the robot also needs to move and operate tools, the architecture also needs to take into account how to combine speech and gestures.

In this work we try to find a solution by leveraging a hierarchical architecture along the lines of CDM and looking more closely into the enablements of interaction as requirements for the dialogue manager, manifested and implemented in a hierarchical architecture.

3 Subsumption Architectures for CDM

The original subsumption architecture shown in Fig. 3 was suggested by Brooks [6] for autonomous robots as a control system based on task achieving behaviours. The proposal abandoned a sequential control model, i.e. a series of functional units which process data from perception to motor control, and instead introduced a layered set of competences that form the control system, a subsumption hierarchy.

The competences are specifications of desired behaviours for a robot in its environment, and a higher level of competence implies a more specific class of desired behaviours. From the processing point of view, information does not flow through the system in a pipeline as in the traditional sequential control models, but in a parallel manner on all hierarchical levels simultaneously. The
The higher-level control system can examine data from the lower-level system, and even suppress or inhibit its normal data flow, thus subsuming the role of the lower level. The lower-level control system is unaware of the level above and operates independently from the higher levels. For instance, to achieve a high-level competence of going to the door, the robot agent can assume competences such as moving around, avoiding obstacles, and building a map of the environment by identifying objects and monitoring changes in the world. Since the higher-level competence includes the lower level one, the higher-level competence can be seen as providing further constraints on it.

3.1 Cascaded Dialogue Architecture

Subsumption architecture is a goal-oriented view of how a robot can operate and learn to act in its environment. It is analogous to the interaction management modelled in CDM where the goal is to produce appropriate dialogue responses. Fig. 4 shows a cascaded architecture for CDM which can be compared with the subsumption architecture in Fig. 3. Dialogue utterances are verbal actions produced by the participants to reach their goals, and thus dialogue management is similar to the robot producing actions as reactions to the changes in its physical environment.

Dialogue management is often implemented as a pipeline where the components for speech recognition, semantic parsing, dialogue state tracking and policy decision, language generation and speech synthesis make up a functional control system (see overviews in [12]). The dialogue task is decomposed differently in the CDM subsumption architecture: the enablements of Contact and Perception are considered desired competences of the agent, subsumed by the competence level of Understanding, which is subsumed by the competence level of Reaction. The levels function in parallel, and each competence level together with the lower-level enablements can make an operational system with the desired competence of that level. For instance, a Perception system, with the lower-level Contact achieves a desired competence of perceiving the environment: the system can detect and recognize communicative signals. However, such a system cannot formulate meaningful messages from the perceived signals; to achieve this competence level, the system must be extended by the desired behaviours of the competence level of Understanding. The higher-level enablements are considered more specific aspects of the interaction, and they subsume the more general enablements associated with the lower levels.

The decomposition of the tasks related to the competences is a major issue in designing and building dialogue management. Decompositions vary depending on the different designs of the implemented systems that follow the needs and requirements of the dialogue domain.

The sets of processing modules correspond to the behavioural competences on a particular level, and the individual modules perform the tasks related to these competences in a rule-based or neural manner. This paper does not go into details of the modules, but Fig.5 shows an example mapping of the CDM layers onto dialogue behaviours that implement various dialogue components. Notice that although CDM can be viewed as a sequential model, the enablements need to operate in a hierarchical way so that desired behaviours are produce by components subsuming the behaviours of the lower levels. The control flow allows parallel processing of the modules as if the architecture were a blackboard-type event-driven architecture.

4 Discussion: Social and situated robots

Dialogue interactions with social robots are "situated": they take place in a dynamically changing world and cannot be totally specified in advance. As the robots can independently move and perceive their environment without explicit human presence (e.g. rescue robots, warehouse robots), the robot’s knowledge of the context can differ from that of the human’s. Moreover, robot agents can also use multimodal signals (e.g. head and gesture movements).
to interact with their partners. Such independent and autonomous observable behaviour makes social robot interfaces qualitatively different from typical computer interfaces like laptops and mobile phones and from chatbots that appear on such devices [9].

The paper has discussed research on the social robot's interactive behaviour modelling within the hierarchical structure of dialogue enables in a subsumption architecture framework. As the tasks and contexts for a robot agent become more complex, requirements for dialogue management also increase. An important issue is to create a joint context in which to ground actions and utterances taking into account both micro- and macro-environments. The context-aware research paradigm emphasizes that AI agents need to be aware of their contexts to support natural, attentive dialogues with humans. Such context-awareness presupposes that the agent is able to recognize the partner's intentions, generate its own intentions, and give feedback on the partner’s actions in the interaction cycle [11].

The methodology consists of developing a computational multidimensional and hierarchical model for interactive behaviour in various interactive contexts. In designing such models, we encounter two main issues that deal with the internal structure of the competence modules: how many different behaviours and behaviour goals are needed for the hierarchy, and are they totally independent or independent of each other. The problem setting is similar to experimenting with different neural models and evaluating their performance with respect to the data set. We plan to experiment with encoding the complexity of the CDM subsumption architecture in different ways, using suitable representations (knowledge graphs) for dialogue competence and content. For this, we will study how to integrate world knowledge as encoded in domain ontologies as well as multimodal signals in the dialogue state representation. Consequently, the structure and management of the dialogue state becomes important.

The robot agent's interaction capability offers a basis for further co-evolution, developing symbiotic relations between robots and humans through learning and interaction. In this respect, interesting questions arise concerning personal attachments to robots. Studies have noted that robots often elicit more personal engagement in human partners due to their less social appearance and behaviour, and often allow more intimate interactions, sharing of personal experiences and “true” opinions than with human partners who may be sensed as having more evaluative attitudes towards the partner’s story. Such personalised systems provide possibilities to tailor responses according to the user, and thus enable long-term interactions with recurring users [16][19]. If the robot’s behaviour becomes more human-like, it is useful to think about dialogue strategies that help the human user to be at ease and conduct trustworthy interactions, but simultaneously be clear that their partner is a robot agent.

The problem is also related to the hypothesis of Boundary Crossing Robots [11][14] and the impact of interactive robot agents on human life and services. Starting development with fairly simple systems and gradually extending their functionalities with the help of users, communities, developers and stakeholders, it is possible that such a spiralling development process which engages a wide range of contributing players will be a way to become more used to robots as companions for every-day tasks.

5 Conclusions
Natural interaction is regarded as a core technology for future society. It is considered to be a facilitator for human-human interaction as well as for communication between humans and smart environments. AI technology is important for creating meaningful services and applications that support humans in their everyday activities.

In this paper we have discussed interaction of human users with situated robots and proposed a new cascaded architecture model based on CDM and subsumption architecture. Future research concerns design and development of innovative conceptual and operational frameworks for modelling human social interactive processes. We plan to apply learning algorithms and inference engines to achieve a high degree of adaptivity of robot agents in a variety of everyday tasks and activities.

Acknowledgments
Thanks go to team members and colleagues for fruitful discussions. The study is based on results obtained from project JPNP20006 commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References


