Towards a Framework for Integrated Natural Language Processing Architectures for Social Robots

Matthias Scheutz[1] and Kathleen Eberhard[2]

[1] Human-Robot Interaction Laboratory, Cognitive Science Program
Indiana University, Bloomington, IN 47401, U.S.A.

[2] Department of Psychology, University of Notre Dame, Notre Dame, IN 46556, U.S.A.

Abstract. Current social robots lack the natural language capacities to be able to interact with humans in natural ways. In this paper, we present results from human experiments intended to isolate spoken interaction types in a search and rescue task and briefly discuss implications for NLP architectures for embodied situated agents.

1 Introduction

Natural language processing on robots is currrently achieved in a sequential fashion, where interpretations are built from utterances in a sequence that mirrors the different linguistic abstractions. Yet, the sequential approch is very different from how humans process language. The implications for human-robot interactions are that robots need a different processing architecture with different processing algorithms if they are to interact with humans in natural language in *natural ways*, for the properties of the human language processing system allows for interaction types that sequential processing systems do not permit. For example, in natural human language interaction, the listener often signals his/her understanding of the speaker's utterance via backchannels or feedback (e.g., head nods, "uh huh" or "mhm"), which overlap with the speaker's utterance at precise points. For such feedback to occur before an utterance has finished, a language processing system must be able to generate partial semantic interpretation on an incomplete sentence. Similarly, human listeners initiate various language-driven actions, from changing eye gaze, to head movements, to gestures and other bodily movements, while the speaker has not finished the utterance. Again, being able to initiate actions based on partial sentences (e.g., looking for referents described by referential phrases) requires a system to be able to determine partial meanings.

In this paper, we will first review some of the psycholinguistic work that analyzes the coordinated goal structure of language interactions as well as the mental processing that underlies rapid incremental comprehension in the face of ambiguity. We then summarize some of the implications for language processing on robots. We will present a human experiment that required two individuals to coordinate with each other via remote audio communication in order to achieve several task goals in a timely fashion.

Scheutz M. and Eberhard K. (2008).
Towards a Framework for Integrated Natural Language Processing Architectures for Social Robots.
In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science*, pages 165-174
DOI: 10.5220/0001741001650174

The results are analyzed with respect to (a) the structure of dialogue that facilitated versus interfered with effective coordination, and (b) the content and form utterances that present challenges to successful comprehension in a remote communicative situation. We conclude with a brief discussion of the implications of our results for NLP architectures for embodied situated agents.

2 Background

Clark (1996) views human language use as a joint project consisting of 4 hierarchical levels of speaker-addressee coordinated actions, which he refers to as an "action ladder". Consider the case of a speaker asking an addressee, "What is the current time?". At the first level, the speaker executes a communicative behavior, which consists of producing the sounds of the utterance. The addressee, in turn, attends to the behavior (speech). At the second level, the speaker presents words and phrases, which are identified as such by the addressee. At the third level, the speaker signals an intended meaning (a request for the current time), and the addressee understands the meaning. At the fourth level, the speaker proposes a joint project, namely that the addressee inform him of the current time, and the addressee considers accepting the proposal. There are two essential properties of this hierarchy of actions. The first is upward causality: The actions at a lower level cause the actions at the next level up. The second property is downward evidence: Evidence of successful completion of the actions at a higher level constitutes evidence of successful completion of the actions at all levels below it.

As Clark (1996, p. 222) states, "A fundamental principle of any intentional action is that people look for evidence that they have done what they intended to do." Furthermore, people strive to provide evidence that is sufficient for current purposes, in a timely manner, and with the least effort. In the example above, valid, timely, and sufficient evidence comes from the addressee responding with the current time soon after the end of the speaker's utterance. In doing so, the addressee provides positive evidence of her acceptance of the speaker's proposed joint project at level 4, as well as positive evidence of her understanding the meaning of the speaker's utterance (level 3), her identification of the speaker's words (level 2), and her attending to the speaker's speech (level 1). In other words, the evidence allows both the speaker and addressee to reach the mutual belief of success at all four levels well enough for current purposes, which is the process of grounding.

Often, a joint project may be extended across a sequence of utterances, as in the case of telling a story, or providing a complex response to a question, or giving a complex direction, where complexity refers to number of propositions or informational units. In these situations, each utterance is an iteration through the first 3 levels, and positive evidence at level 3 (understanding) is provided by addressees in the form of acknowledgments, which may be verbal (e.g., yes, uh huh, mkay, okay) or nonverbal (head nods). Acknowledgments may occur on a separate turn or they may overlap with the speaker's utterance (i.e., Yngve's (1970) backchannels).

Most psycholinguistic research as well as research in developing artificial natural language processing systems has focused on the processes involved at the first 3 levels of action (i.e., producing and perceiving a speech signal, identifying words and the

phrase structure, understanding the propositional content of the utterance, including the establishment of referents). The former research has shown that humans rapidly integrate bottom-up constraints from the linguistic input, such as the sounds, words, syntax, and semantics with top-down constraints from the discourse and pragmatic context, such as the set of possible referents (e.g., [1]) and expectations about the speaker's communicative goals (e.g., [2]). The strength of both the linguistic and contextual constraints depends on their availability. The availability of linguistic constraints is affected by the clarity of the acoustical signal, the fluency and rate of speech, the frequency of the words, the specificity of their meaning, and the frequency, complexity and specificity of syntactic structure. The availability of contextual constraints is determined by the "quality of evidence" for the bases of the speaker's and addressee's common ground [3] or shared knowledge. The quality is high when, among other things, both the speaker and addressee know the goal structure of the communicative task, and the set of relevant referents is visually co-present (e.g., [4]. Importantly, the availability of linguistic constraints interacts with the availability of contextual constraints in the incremental construction of an interpretation [5], such that when the linguistic constraints are weak or underspecified there will be greater reliance on contextual constraints [6]. Furthermore, speakers are likely to produce weakly underspecified utterances when there are strong contextual constraints for their interpretation (i.e., when there is reliable evidence for the speaker's and addressee's common ground). For instance, in face-to-face conversations about visually co-present referents, speakers may use short deictic expressions accompanied by indicative gestures (e.g., saying, "move the box over there", while pointing to the location that is the referent of "there") [7, 4].

3 Experiment and Results

To be able to isolate the design principles that are required for an NLP system for robots that interact with humans in natural ways, we designed a human experiment in which two individuals must coordinate with each other via remote audio communication to accomplish several tasks. In particular, one person, the "director", direct the other person, the "member", through an unfamiliar environment to locate and perform various actions on target objects scattered throughout the environment. In the following we will first describe the experimental task, and the report some of the findings that are useful for extracting principles of the processing architecture.

The chosen task is a team search task where two humans, who are not co-located in the same physical space, must coordinate their actions using natural language to accomplish several goals within a limited amount of time. One individual is assigned the role as the director, the other is assigned the role of member. Neither was familiar with the search environment, which consisted of several (cluttered) rooms and a surrounding hallway. The director was seated at a table in a quiet room outside of the environment. S/he wore headphones and a microphone for communicating with the member. The member wore a helmet fitted with a camera for recording the visual scene (not viewable to the director), and a microphone and headphones for communicating with the director.

3.1 Procedure

At the beginning of the experiment, both the director and member were told that the director would be given a map of the search environment, which consisted of all rooms with open doors. The director was told that s/he would have to remain seated at a table in a room that was external to the search environment. They were told that the director's map showed the locations of a cardboard box, a number of blue boxes containing colored wooden blocks, and 8 empty pink boxes. They were also informed that the lab environment contained 8 empty green boxes, which were not shown on the map. They were told that there were several tasks that needed to be completed as quickly as possible:

1. The member was to tell the leader the location of each of the 8 green boxes, which had numbers written on them. The leader was to mark the map with the location of the green boxes by writing their number on the map.
2. The leader was to direct the member through the environment to the location of the cardboard box, which the member was to retrieve.
3. The member was to then empty the blocks in all of the blue boxes into the cardboard box, leaving the blue boxes in their location. The leader was to assist the member with finding the blue boxes by giving directions to them from the map. However, they were informed that some of the locations of the blue boxes on the map would be inaccurate, and that the map did not show the location of all of the blue boxes.
4. They were told that instructions for the pink boxes would be given at some point during the task.

The director and member were told that each would receive $5.00 for participating in the experiment and that each would receive an extra $5.00 if they successfully completed all of the tasks. After a sound check, the member began walking through the environment.

After 5 minutes, the experimenter interrupted the director and informed him of the task for the pink boxes: Each of the blue boxes contained a yellow block, and the member was to place one yellow block into each of the 8 pink boxes. In addition, the team had only 3 minutes left in which to complete all of the tasks (recording the location of the green boxes on the map, emptying the blocks from the 8 blue boxes into the cardboard box, and putting one yellow block into each of the 8 pink boxes). A cooking timer with an audible ticking sound was set to 3 minutes and placed on the table in front of the director. Then the experimenter left the room. The experiment ended when the bell on the timer rang.

3.2 Results

There were 7 pairs of subjects run in the experiment. The first pair was eliminated because of problems with the audio recording equipment at the beginning of the experiment. The second pair was eliminated because of poor audio recording. Thus, data were collected for the 5 remaining pairs.

The results in Table 1 show that, with the exception of Team #5, there is no relation between the number of green and blue box tasks completed during the 1st 5 minutes and the grand total at the end.

Table 1. Table showing the number of tasks involving the green, blue, and pink boxes that were successfully completed by each team. The maximum number for each type of box is 8. The teams are sorted according to the total number of tasks completed. *Team #5 was the only team that did not retrieve the cardboard box (for collecting the blocks from the blue boxes) before the 3 minute warning.

	1st 5 min.			Last 3 min.					
Team	green	blue*	green+blue	green	blue	pink	Tl green	Tl blue	TOTAL
7	4	6	10	3	2	8	7	8	23
4	8	1	9	-	6	6	8	7	21
6	6	2	8	0	4	6	6	6	18
3	8	2	10	-	3	2	8	5	15
5	7	0	7	1	2	2	8	2	12

Dialogue Structure. The interactions between the pair that was most successful in completing the task goals (Team 7) were compared with the interactions between the pair that was least successful (Team 5) in order to identify structures of dialogue that characterize effective vs. ineffective coordination, respectively.[3]

For all teams, at the beginning of the experiment the nature of the task resulted in the overarching goal in which the director uses his/her map to direct the member through the multi-room environment to the location of the cardboard box, and an embedded goal, in which the member reports the location of each green box as it is encountered along the way. Thus, following Clark's (1996) 4-level "action ladder" framework, the overarching goal was an extended joint project requiring a sequence of directive utterances that were subordinate joint projects, and the grounding of which required the director and member to reach mutual belief of the member's location in the environment.

Team 7's dialogue begins with the director (D) proposing the overarching goal and the member (M) acceptance of it:

Example from Team #7:

```
1 D: from this first hole do you wanna get the cardboard box?
2 M: yes
3 D: alright let's do it
```

The embedded goal resulted in multiple individual joint projects, the grounding of which required the director and member to reach mutual belief that the director sufficiently marked the location of a green box on his/her map. One aspect of Team 7's dialogue that made it effective was that the addressee routinely provided evidence of understanding the speaker's direction as well as evidence of when the directed action (joint project) was completed. This routine may have been encouraged by the director's explicit request for the latter evidence during the exchange that constituted the first sub joint project of the overarching goal:

[3] Note that all subsquent transcriptions are verbatim and include disfluencies (false starts, repairs, etc.). Pauses are indicated by periods and syllable lengthenings, such as pronouncing "the" as "thee", are indicated by a colon following the vowel that is lengthened, e.g., "the:".

Example from Team #7:

```
4 D: um . go straight through the room you're in to an open door
     that's right across from you
5 M: alright
6 D: let me know when you get there
7 M: I'm at-I'm at the open door
```

So, in line 5, the Member's response acknowledges understanding and acceptance of the director's direction (sub joint project). In line 6, the director explicitly requests verbal evidence from the member of the completion of the directed movement. The member does provides this evidence in line 7, and, furthermore, his description of his location provides more reliable evidence than simply providing evidence in the form of acknowledgment such as "okay".

The sequence above continues below, with an embedded joint project in which the member describes the location of a green box that is to be marked by director on the map. Like the exchange above, but with the director and member's roles reversed, the director provides evidence of understanding the member's description, by repeating it, followed by an acknowledgment by the member, and then the director providing evidence of the action's completion:

Example from Team #7:

```
8 D: okay go through the open door and towards the steps that are
     right in front of you before the steps take a . take a right
9 M: okay . uh right-right on the steps there's a green box
     number two
10 D: oh number two right on the steps
11 M: yeah
12 D: okay I got it
```

Note that the Member's utterance in line 9 simultaneous proposes an embedded joint project (marking the location of a green box on the map) but also provides evidence of the completion of the sub joint project proposed by the director's utterance in line 8. That is, upon completion of the embedded joint project, the director can assume that the member's location is near the steps. Thus, the director's proposal of the next sub joint project begins with that assumption:

Example from Team #7:

```
13 D: alright . if you're looking at the steps you take a right
      there should be another open door
14 M: so don't actually go up the steps
15 D: don't actually go up the steps
16 M: okay
17 M: yep I see the door
```

The exchange above also contains a side sequence that is initiated by the Member's request for clarification in line 14. The side sequence ends with the member's acknowledgment in line 16 of the director's clarification in line 15, and the sub joint project is completed with the member's description of the door in line 17.

Unlike Team 7, Team 5's dialogue lacks orderliness in providing evidence of both understanding a proposed joint project and its completion. This is illustrated in the first two lines of the transcript below, where the member neither requests nor receives evidence from the director of his understanding and acceptance of her description of the

location of the third green box in line 33 (i.e., an embedded joint project). More importantly, however, as the complete exchange below shows, there was no established agreement between the director and member on achieving the overarching goal of locating and retrieving the cardboard box by having the director direct the member to the box's location. Specifically, this agreement does not occur until line 47 in the transcription below, which occurs approximately 3.5 minutes into the task. Until that point, the member simply provided descriptions of her movement through the environment along with descriptions of the location of green boxes (the embedded goal/joint project).

Example from Team #5:

```
33 M: Okay I'm going forward and then taking a right and the first
      bo-in the f-to the first room there . so: right now  I've got
      um a green box number three on the chair on my right
      it's-it-as soon as I'm in the doorway I'm facing forward
      green box on my right
34 M: and there's also a blue box on my right but I don't have the
      brown box yet so I'm gonna turn around and keep looking for
      the brown-or go back and look for the brown box or somethin
35 D: alright the brown box-let me-alright t-two questions for you
      now you said you walked in u:h you walked in that- that room
      which is . to the right of the doorway
36 M: uh huh
37 D: now you said as soon as you walked in, there was a chair on
      your right hand side?
38 M: yep
39 D: with it-so it's basically on the wall where the door is
40 M: it's a little bit off the wall but it's like maybe my foots
      worth of a distance between the
41 D: mkay
42 D: alright and that's number three
43 M: yes
44 D: alright there's a blue box in that room
45 M: yes
46 D: and you said that-uh the cardboard box is-is basically all
      the way to the end so should I-should I se-should I send
      you-do you want me to send you to where the cardboard box is
      and then we can backtrack
47 M: u:m yeah that's fine
```

Having established the overarching goal as well as established mutual belief as to the member's current location in the environment, the exchange above continues with the director proposing the first sub joint project, followed by the second. The member provides evidence for both in the form of an acknowledgment, which is ambiguous with respect to being evidence for level 3 (understanding the direction) or level 4 (acceptance and completion of the directed action). This ambiguity results in a side sequence in which the director requests clarification with respect to the member's current location:

Example from Team #5:

```
48 D: alright . uh so you're gonna wanna go back-step back out of
      the room you were just in
49 M: uh huh
50 D: and continue in through that doorway that was on your left
51 M: uh huh
52 D: u:h . are you in that room already?
53 M: yep
```

At this point, which is slightly more than 4 minutes into the task, the member proposes explicitly establishing of agreement on the embedded goal of her informing the director of the location of the green boxes.

Example from Team #5:

```
54 M: I actually see like three more green boxes do you want
      those now or do you want those later
55 D: uh whatever you wanna do we can stop and get those on
      the way if you want
56 M: okay let's just do that now
```

The member continues with proposing an embedded joint project (description of the location of a green box); however, the complexity of her proposal results in a side sequence initiated by the director who needs further clarification before accepting/completing the embedded joint project:

Example from Team #5:

```
57 M: so as I walked in I go to the room on my right there's two
      filing cabinets on the second filing cabinet there's box
      number four green box number four
58 D: alright as-the second one close to the wall that you are
      entering in
59 M: I walk in there's one on my right and then there-I just
      make another step it's-it's the second one on my right so
      this is gonna be the second one
60 D: alright and that's number four you said?
61 M: yes
62 D: okay
```

3.3 Content and Forms of Utterances

As the example transcripts above illustrate, disfluencies were the norm, not the exception, which are potential impediments to the director and member's successful grounding of understanding. The disfluencies include frequent pauses within utterances often signalled with the fillers "um" or "uh". In addition, there are numerous repairs (e.g., "the first bo-in the f-to the first room"), false starts/repetitions (e.g., "it's-it's the second one on my right", "uh right-right on the steps", "so should I-should I se-should I send you do-you want me to send you"), and omissions of words (e.g., "I'm facing forward green box on my right"), which result in ungrammatical utterances. There also are instances of uncorrected speech errors such as substituting the words "block" and "book" for the intended word "box".

In addition, there are numerous examples of lexically ambiguous words, most notably, the word, "right", which often would occur several times within a sequence of utterances, with each occurrence corresponding to a different meaning (i.e., an acknowledgment (correct), a direction (vs. left), and an intensifier (right there)).

Example from Team #5:

```
M: alright . okay I'm going forward and then taking a right and
   the first bo-in the f-to the first room . there so: right now
   I've got um a green box number three on the chair on my right
   it's-it-as soon as I'm in the doorway I'm facing forward green
   box on my right.
```

Much of the ambiguity in the linguistic input can be resolved by the contextual and pragmatic constraints resulting from the director and member's shared knowledge of the task's goals and their shared knowledge of the environment and referents in it that is provided by the correspondence between the director's map and physical environment.

3.4 Implications for NLP Architectures for Natural spoken Language

The main implications for designing a natural language processing architecture for robots is that, different from the standard picture of constructing meanings out of sentences, meanings are obtained from interactions that serve particular purposes and accomplish particular goals. Language here serves a coordinating role in establishing a joint project and humans define those projects, agree on them, and keep track of them until they are accomplished or the goal structure changes. The goal structure can also be seen as imposing constraints on the natural language processing system that allows for dealing with disfluencies and ambiguity of various kinds. Moreover, perception, action, and language processing are all instrinsically intertwined, sometimes involving complex patterns of actions, utterances and responses, where meaningful linguistic fragments result from their context together with prosodic, temporal, task and goal information, and not sentence boundaries. An NLP architecture for robots, therefore, needs to be able to process language in the same kind of interactive, goal-oriented way that humans use; this includes the timing of utterances, non-linguistic information, backchannel feedback, and any other component involved in establishing meaning (for a first step towards implementing some of these principles, see [8–12]).

4 Conclusions

In this paper, we argued that processing multiple linguistic, perceptual, and contextual constraints incrementally and determining partial meanings to be able to provide backchanneling feedback and initiate actions early is of critical importance for robots that are supposed to interact with humans in natural language in natural ways. We reported results from human experiments in a search task that demonstrated these and other important principles that can be used for specifying a natural language processing architecture for robots which will allow robots to engage in more human-like interaction patterns.

Acknowledgements

This work was in part funded by an ONR MURI grant N00014-07-1-1049 to both authors.

References

1. Chambers, C., Tanenhaus, M., Magnuson, J.: Actions and affordances in syntactic ambiguity resolution. Journal of Experimental Psychology: Learning, Memory, and Cognition 30 (2004) 687–696
2. Hanna, J., Tanenhaus, M.: Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. Cognitive Science 28 (2004) 75–88
3. Clark, H.H.: Using Language. Cambridge University Press, Cambridge (1996)
4. Clark, H., Krych, M.: Speaking while monitoring addressees for understanding. Journal of Memory and Language 50 (2004) 62–81

5. Tanenhaus, M., Trueswell, J.: Sentence comprehension. In Miller, J.L., Eimas, P.D., eds.: Handbook of Perception and Cognition. Vol 11: Speech, Language, and Communication. Academic Press, Orlando (1995) 217–262
6. Reason, J.: Human Error. Cambridge University Press, New York (1990)
7. Bangerter, A.: Using pointing and describing to achieve joint focus of attention in dialogue. Psychological Science 15 (2004) 415–419
8. Brick, T., Scheutz, M.: Incremental natural language processing for hri. In: Proceedings of the Second ACM IEEE International Conference on Human-Robot Interaction, Washington D.C. (2007) 263–270
9. Brick, T., Schermerhorn, P., Scheutz, M.: Speech and action: Integration of action and language for mobile robots. In: Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA (2007) forthcoming
10. Scheutz, M., Schermerhorn, P., Kramer, J., Anderson, D.: First steps toward natural human-like HRI. Autonomous Robots 22 (2007) 411–423
11. Scheutz, M., Schermerhorn, P., Kramer, J., Middendorff, C.: The utility of affect expression in natural language interactions in joint human-robot tasks. In: Proceedings of the 1st ACM International Conference on Human-Robot Interaction. (2006) 226–233
12. Scheutz, M., Eberhard, K., Andronache, V.: A real-time robotic model of human reference resolution using visual constraints. Connection Science Journal 16 (2004) 145–167