Felix Gervits

Tufts University
Human-Robot Interaction Lab
200 Boston Ave.
Medford, MA 02155

Felix.Gervits@tufts.edu

1 Research Interests

My primary research goal is applying models of human communication and dialogue to robotic systems in order to improve coordination in multi-agent human-robot teams..

One of the central challenges in this endeavor is the need to handle fast-paced natural language dialogue in performance settings. This is difficult because task-based dialogue is often very messy, riddled with disfluency, agrammatism, overlapping speech and ambiguity. Another challenge involves monitoring the mental state of one's teammates and performance of the team as a whole (Sycara and Sukthankar, 2006). The robot will need to monitor team communication channels to identify changes in performance state and team cohesion, and then modify its behavior to repair any problems.

There is a wealth of human factors literature on team performance that addresses how human teams overcome some of these challenges, and my research utilizes this knowledge base to inform the design of **dialogue systems for human-robot interaction (HRI)**. I also use methods from a variety of disciplines, including **corpus-based discourse analysis**, **team performance evaluation**, and **integrated architectures for NLU**.

1.1 Task-based dialogue in humans

My research has so far consisted of a systematic investigation of task-based dialogue in humans. The data was obtained from the Cooperative Remote Search Task (CReST) corpus (Eberhard et al., 2010), which contains hours of collaborative dialogue between remotely-communicating partners performing a joint-task. Since CReST was designed to simulate the structure of teams in which a robot may play a vital role (e.g., search and rescue missions), the results of this analysis inform our understanding of the kinds of communication and coordination strategies that artificial agents will need to adopt to be effective partners in these mixed-initiative teams.

To get at the linguistic- and dialogue-level properties in the corpus, I wrote a program to parse and filter out various features from the transcribed text data. These features include 5 types of disfluencies, pauses, speech rate, average utterance length, and dialogue moves. I then performed statistical analyses on these features to establish which if any were correlated with team effectiveness (Gervits et al., 2016).

Interestingly, I found that disfluency rate (particularly self-repairs) *increased* for the effective teams. This seemingly counterintuitive result suggests that disfluencies are not only caused by production difficulty, but rather can serve as collaborative tools in the discourse to enhance coordination and performance. Another novel finding showed that the best performing teams were those that more efficiently grounded their conversational exchanges and minimized joint collaborative effort through particular discourse patterns - namely: Check and Ready dialogue moves, frequent acknowledgments, and establishing shared referents.

These results have important implications for improving coordination in NLU systems and in mixed-initiative human-robot teams. The finding that self-repairs are strong indicators of collaborative process and are increasingly utilized in effective teams, suggests that the detection and identification of speech disfluency is crucial for robust NLU. For example, speech rate could signal increasing workload demands (Berthold and Jameson, 1999), self-repairs could indicate grounding, and the receptiveness of a speaker to monitor their teammate (Clark and Krych, 2004), and fillers can provide clues about turn-taking and discourse structure (Swerts, 1998). It is important that dialogue systems are able to utilize this information contained in disfluent utterances to gain insight into speakers' meta-cognitive states.

1.2 Integrated HRI architecture

The central focus of my research going forward involves implementing these markers of team effectiveness in the DIARC architecture (Schermerhorn et al., 2006). DIARC is an integrated cognitive-robotic architecture that also serves as a platform and test bed for HRI experimentation. It is a robust system which integrates vision, motor control, goal planning/action, as well as advanced natural language processing. On the language side, the system is capable of speech recognition, incremental parsing, reference resolution, and context-sensitive pragmatic reasoning.

My ongoing involvement with the architecture has been to develop the dialogue manager in order to handle more interactive, dynamic exchanges - such as those found in the CReST corpus. I am working on several concurrent projects:

- Handling disfluent utterances. This presents a major challenge for current NLU systems, which generally cannot parse disfluent utterances. I am currently developing algorithms to detect different types of disfluencies, including self-repairs, filled pauses, and lexical fillers (“like”, “you know”), and to enable the system to utilize information contained in the disfluencies to improve speech recognition.
- Contextual modulation to determine if the literal or nonliteral interpretation on an utterance is to be used. For example, the utterance “Can you walk forward?” could be interpreted literally as a question about the robot’s capabilities, or nonliterally as an indirect command to walk forward. The goal of this project is to develop algorithms to automatically detect which meaning is intended based on contextual cues.

2 Future of Spoken Dialog Research

The ultimate goal of spoken dialogue research is to develop robust systems that can interact with humans in natural, real-world settings. Though we are currently nowhere near achieving this goal, I believe that we are furthest behind in the areas of 1) situated interaction, 2) extra-linguistic communication, and 3) disfluency handling. Given that human interaction largely takes place between embodied agents co-located in a shared physical environment, it is important that dialogue systems can make use of the shared perceptual context available in such settings. Additionally, dialogue systems should be able to handle facial expressions, gesture, back channel feedback, and many other non-verbal cues, as these are ubiquitous in natural human interactions. Finally, efforts at speech recognition and NLU need to be directed towards handling more natural kinds of utterances, which include disfluencies and agrammatical phrases of all sorts. Parsing should be incremental in order to respect human timing, and the system should have ways to handle errors and repairs in a seamless way in order to preserve the interaction if something goes wrong.

3 Suggestions for Discussion

- To what extent do design decisions for spoken dialogue systems need to be grounded in the empirical literature? What are the pros and cons of this approach?
- What are some ways to integrate non-verbal communicative signals from various modalities (facial expressions, gaze location, gestures, etc.) into the interpretation of an utterance? What are some existing systems that do this?
- How can ”messy” features of speech (i.e., disfluencies) be detected and utilized by dialogue systems to improve performance.

References

Berthold, Andre, and Jameson, Anthony. 1999. *Interpreting Symptoms of Cognitive Load in Speech Input*. User Modeling: Proceedings of the Seventh International Conference, pp. 235-244.

Clark, Herbert H. and Krych, Meredyth A. 2004. *Speaking while monitoring addressees for understanding*. Journal of Memory and Language, 50(1), pp. 62-81.

Eberhard, Kathleen, Nicholson, Hannele, Keubler, Sandra, Gudersen, Susan, and Scheutz, Matthias. 2010. *The Indiana Cooperative Remote Search Task (CReST) Corpus.* Proceedings of the International Conference on Language Resource and Evaluation, (LREC) 2010, 17-23.

Gervits, Felix, Eberhard, Kathleen, and Scheutz, Matthias. 2016. *Team communication as a collaborative process*. Manuscript submitted for publication.

Schermerhorn, Paul, Kramer, James, Brick, Timothy, Anderson, David, Dingler, Aaron, and Scheutz, Matthias. 2006. *DIARC: A testbed for natural human-robot interactions*. Proceedings of AAAI 2006 Mobile Robot Workshop.

Swerts, Marc. 1998 *Filled pauses as markers of discourse structure*. Journal of Pragmatics, 30(4), pp. 485-496.

Sycara, K. and Sukthankar, G. 2006. *Literature review of teamwork models*. Technical report CMU-RI-TR-06-50.

Biographical Sketch

Felix is a PhD student at the Human-Robot Interaction Lab at Tufts University advised by Dr. Matthias Scheutz. He has a Bachelor’s degree in Cognitive Science from Rensselaer Polytechnic Institute and a Master’s degree in Linguistics and Cognitive Science from the University of Delaware. Felix has a diverse multidisciplinary research background, with experience in the fields of AI, Cognitive Neuroscience, and Psycholinguistics. He is currently using this diverse background to inform the design of robust, spoken dialogue systems for human-robot interaction.