Artificial Intelligence for Human-Robot Interaction
Papers from the AAAI 2015 Fall Symposium

The RoboHelper Project: From Multimodal Corpus to Embodiment on a Robot

Barbara Di Eugenio and **Miloš Žefran**
Computer Science / Electrical and Computer Engineering
University of Illinois at Chicago
Chicago, IL 60607

Abstract

In this position paper, we describe the RoboHelper project, its findings and our vision for its future. The long-term goal of RoboHelper is to develop assistive robots for the elderly. The main thesis of our work is that such robots must crucially be able to participate in multimodal dialogues.

Contributions of our work to date include the ELDERLY-AT-HOME corpus that we collected and annotated. It consists of 20 task-oriented human-human dialogues between a helper and an elderly person in a fully functional apartment. The unique feature of the corpus is that in addition to video and audio, it includes recordings of physical interaction. Based on this data, we have demonstrated the crucial role that Haptic-Ostensive (H-O) actions play in interpreting language and uncovering a person's intentions. H-O actions manipulate objects, but they also often perform a referring function. Our models were derived on the basis of manually annotated categories. Additional experiments show that we can identify H-O actions using the physical interaction data measured through an unobtrusive sensory glove developed as part of the project.

In future work, we will derive models for the robot to decide what to do next (as opposed to interpreting what the interlocutor did); explore other types of physical interactions; and refine preliminary implementations of our models on the Nao robotic platform.

The Elderly-at-Home Corpus and H-O Actions

Human-human collaborative dialogues are inherently multimodal. This implies that for Human-Robot Interaction (HRI) to be perceived by humans as natural, it should incorporate multimodality as well. In addition to speech, a robot needs to be able to interpret and generate different kinds of gestures as well as actions that involve physical contact. Besides making the interaction natural, multimodal dialogues have other advantages (Cohen and McGee 2004; Jaimes and Sebe 2007): multimodality makes the dialogue more robust and more efficient. Whereas pointing gestures have been broadly studied, *Haptic-Ostensive (H-O) actions* (Foster et al. 2008) have not. H-O actions are actions involving physical contact that manipulate objects and may simultaneously refer to them.

The three main contributions of the RoboHelper project so far are summarized below - further information can be found in (Chen and Di Eugenio 2012; Chen and Di Eugenio 2013; Chen 2014; Javaid, Žefran, and Di Eugenio 2014; Javaid 2014; Chen et al. 2015).

Corpus Collection and Analysis. The ELDERLY-AT-HOME corpus of human-human collaborative dialogues was collected in a fully functional studio apartment at Rush University in Chicago. Each experiment involved interaction between an elderly person (ELD) and a helper (HEL). The experiments focused on four Activities of Daily Living (ADLs) (Krapp 2002): walking; getting up from a bed or a chair; putting on shoes; and preparing dinner. Helpers were two gerontological nursing students. The corpus is composed of five pilot experiments (two faculty members played the role of ELD); and of fifteen experiments with a real elderly person, who resided in an assisted living facility and was transported to the apartment mentioned above. All elderly subjects were highly functioning at a cognitive level and did not have any major physical impairments. Figure 1 shows a video frame from the experiment recordings. During the experiments, both participants were wearing a microphone, and a sensory glove on their dominant hand to collect physical interaction data.

Figure 1: Data Collection Experiments

An experiment usually lasted about 50 minutes, but that included irrelevant content (obtaining informed consent, interventions from the project staff); hence, on average we obtained about 15 minutes of what we call *effective* data, for a total of 4782 spoken utterances. All the data was transcribed. We then focused on *Find* tasks, continuous time spans dur-

ELD: Um, so can we start maybe with taking **the pot** out, um
[*DA=Instruct*]
ELD: And then you can help me fill it with water maybe?
[*DA=Instruct*]
HEL: Of course
[*DA=Acknowledge*]
HEL: Okay, so the pots and pans are down here?
[*DA=Query-yn; H-O action=Touch, target=Cabinet6; H-O Action=Open, target=Cabinet6*]
ELD: yes, there are, but maybe not the one I want to use
[*DA=state-n*]
ELD: No, we need a bigger one.
[*DA=reply-n*]
ELD: Look down **there**
[*(DA= Instruct; point, target= Cabinet7*]
HEL: Okay
[*DA=Acknowledge, H-O Action=Close, target=Cabinet6)*]
HEL: [*H-O Action=Open, target=Cabinet7*]
ELD: Yeah, **that one**
[*DA=state-y*]
HEL: **This one**?
[*DA=check; H-O action=Touch, target=Pot2*]
ELD: Yes
[*DA=reply-y*]

Figure 2: An excerpt from one of the RoboHelper dialogues

ing which the two subjects are collaborating on finding objects, such as dinnerware, pots, food, etc.

The *Find* subcorpus was annotated for referring expressions and their antecedents, dialogue acts, physical interaction actions, including H-O actions, and pointing gestures. For the latter two, their target is also annotated. The corpus was annotated by means of Anvil (Kipp 2001), a multimodal annotation system.

Consider the exchange in Figure 2 (HEL is the HELper and ELD the ELDerly person). In boldface, we show referring expressions; in italic, some annotations. DA (Dialogue Act) captures the main intention of the speaker. We defined 13 dialogue acts, adding three new dialogue acts to the original inventory of twelve dialogue acts in the MapTask corpus (Anderson et al. 1991). [1]. Whereas many other DA coding schemes exist (for example (Levin et al. 1998; Graff, Canavan, and Zipperlen 1998; Shriberg et al. 2004)), we chose MapTask because it is simple but expressive, and targeted for task-oriented collaborative tasks. However, neither MapTask nor many other coding schemes include DAs that apply to utterances that are used to respond to gestures and actions, such as *Yeah, that one* in Figure 2. Hence, we added three DAs: (*state-y*, a statement which conveys "yes", such as *Yeah, that one* in Figure 2; **state-n**, a statement which conveys "no", e.g., *yes, there are, but maybe not the one I want to use* in Figure 2; **state**, still a statement, but not conveying acceptance or rejection, e.g., *So we got the soup.*

We annotate for two types of extra-linguistic information: pointing gestures and H-O actions. Both of them capture hand gestures that indicate their targets; the distinction is that for a pointing gesture, there is no physical contact between the hand and the target(s). On the contrary, H-O actions by definition involve physical contact with the manipulated objects. Since in our corpus head movements, or other body part movements, seldom if ever indicate targets, they are excluded from our pointing gesture annotation.

Two attributes are marked for a pointing gesture: time span and target. To mark the physical targets of pointing gestures, we devised a referring index system that assigned a static index to objects with fixed locations, like cabinets, drawers, fridges, etc. (Cabinet 6 and 7 in Figure 2). Movable objects which have many different instances, such as cups, glasses, etc. are assigned a run-time referring index (e.g. Pot2 in Figure 2), which means this is the second pot appearing in the experiment. The same indices are used for the references of H-O actions.

Since no prior work defines types of H-O actions, we defined five of them. They resulted from the following two observations: (1) these five H-O types are empirically based on our data, namely, these H-O actions are frequently observed in the corpus. (2) They are within the scope of what we envisioned could be recognized from the signals measured by the pressure sensors on the sensory gloves.

- *Touch*: The subject touches the target, but no immediate further action is performed.
- *Grasp-Show*: One subject takes out or picks up an object, holds it stably for a short period of time, and (in the judgment of the annotator) intentionally shows it to the other subject.
- *Grasp-No-Show*: Still a grasp, but (in the judgment of the annotator) the subject does not intentionally show it to the other subject.
- *Open*: Starts when the subject makes physical contact with the handle of the fridge, a cabinet or a drawer, and starts to pull; ends when the physical contact is off.
- *Close*: Starts when the subject has physical contact with the handle of the fridge, a cabinet or a drawer, and starts to push; it ends when the physical contact is off.

We believe ours is the first extensive corpus analysis of H-O actions. In addition to H-O actions, where the target of the action is an object, we also defined, and annotated for, *physical interactions*: direct or indirect force exchanges between two subjects. A direct force exchange occurs when one subject's hand is on a body part of the other subject; an indirect one occurs when two subjects exchange forces through a third object like: bowl, pot, spoon, silverware, etc. As for H-O actions, no inventory of Physical Interactions exist, hence we defined our own. They include *Provide Support* (as when HEL helps ELD to stand up); the pair *Give/Receive*, which occurs when a subject is giving an object to the other subject, who is hence receiving it; and *Hold*, when two subjects are holding or carrying objects together.

As of the time of writing, we are putting the last touches on making the ELDERLY-AT-HOME publicly available; we believe it will be of great use to the community.

Multimodal Dialogue Processing. On the basis of our corpus and annotations, we developed the reference resolution module, and the dialogue act classifier. In both components, H-O actions and extra-linguistic information in general, play

[1] http://groups.inf.ed.ac.uk/maptask/interface/expl.html#moves

a crucial role. Our referential resolution module focuses on third person pronouns (e.g. *it*) and deictics (e.g. *there, this*). It comprises standard components to decide whether PR (third person pronoun or deictic) is referential (an example of non referential *there* can be found in in *yes, there are, but maybe not the one I want to use* in Figure 2). A second module then generates all potential coreference candidate pairs for PR, each of which comprises the referential expression PR in question and one potential antecedent; crucially, potential antecedents can be objects that were introduced extra-linguistically in the context, via a pointing gesture or an H-O action. The core of the reference resolution module is the candidate pair classifier, that labels each candidate pair as true or false. We developed the candidate pair classification model via the following machine learning algorithms: MaxEnt, Decision Tree, and Support Vector Machine (SVM). Features include distance (measured in seconds or with respect to previous moves in the interaction), various types of agreement, and H-O type if the antecedent had been introduced by such an action. If antecedents introduced extra-linguistically are not used in the candidate pair generation phase, only a mere 4% of the candidate expressions can be resolved. Using the targets of both pointing gestures and H-O actions as potential antecedents results in an accuracy of 47.4%; H-O actions account for 50% more correctly resolved pronouns and deictics with respect to only using pointing gestures. The algorithm that performs best is MaxEnt.

We approached the recognition of dialogue acts via machine learning models as well, in this case experimenting with MaxEnt, Naive Bayes and Decision Tree. Also for this task, we showed that extra-linguistic information significantly improves accuracy with respect to models that don't include this information. Specifically, when adding a triad of extra-linguistic features (pointing gestures, H-O actions, and location of the speaker) to linguistic features, accuracy significantly improves from .641 to .666. The best result (.746) is obtained when additional information, such as dialogue history, is further added. All these accuracies come from MaxEnt, which outperforms the other two algorithms by at least 10 percentage points. Contrary to the reference resolution case, we were not able to assess the independent contributions of pointing gestures and H-O actions.

Recognition of H-O actions from the physical interaction data. The models of referring expressions and dialogue acts just mentioned are based on H-O actions *annotated by a human*. To ground our corpus analysis and modeling in the physical world, we performed further data collection to study whether (some) H-O actions can be automatically recognized from the physical interaction data collected through the sensory glove instrumented with pressure sensors and an inertial navigation unit (IMU). Dynamic Time Warping (Sakoe and Chiba 1978) was used for the H-O recognition, with within-subject recognition rate between .588 (when the training data was obtained in a different experimental setup than the test data) and .948 (when the training and test data came from the same data set). These experiments show that the H-O actions of interest can be recognized within subjects, even though pressure sensors are relatively imprecise and the data provided by the sensory glove is noisy.

Current and Future Work

Our work has several direct implications for HRI. One of the more challenging aspects of HRI is how to manage physical interaction between the robot and the human. Traditionally, data from human-human interactions is used to formulate successful strategies for the robot. However, it is difficult to collect physical interaction data unobtrusively. Through the RoboHelper project we demonstrated that pressure sensors can provide useful information about the physical interaction actions that are part of a multimodal interaction. Next, we showed that H-O actions, a subset of physical interaction actions, play a crucial role in reference resolution and dialogue act classification. In turn, the ELDERLY-AT-HOME corpus is one of the few multimodal corpora, if not the only one, that contains physical interaction data of the sort we describe. Finally, the algorithms we developed for automatic recognition of H-O actions from the physical interaction data can be directly used to expand the ELDERLY-AT-HOME corpus as new data is collected, and can provide the input for learning by demonstration in the assistive robotics domain.

However, several issues need to be resolved to make the work fully usable from an HRI perspective. From a dialogue management point of view, we have not addressed the problem of planning the next action; we will develop a module that would decide what to do next, and generate a spoken utterance and / or a pointing gesture / H-O action. We can partially address *what to do next* at a purely symbolic level, by applying machine learning to this task too, on the sequences of turns that occur in our corpus. However, this module should clearly be informed by feedback from the robot as well, since tasks that are considered primitive at the dialogue level may require several physical actions and thus result in further decomposition. For example, *Can you see if you find a fork* results in HEL moving towards the cabinets before opening any of them. We also need to expand our methodology to include the context into both recognition of the current dialogue state as well as action planning. This will require collecting additional data and building a hierarchy of models all the way to the task-level.

While we have successfully demonstrated H-O action recognition from the physical interaction data within subjects, the developed algorithms need to be generalized and made more robust to be really useful for HRI. This involves both improving the developed data glove as well as exploring alternatives to pressure sensors such as artificial skin (Noda et al. 2007; Silvera-Tawil, Rye, and Velonaki 2014; Barron-Gonzalez and Prescott 2014). Further, we have demonstrated the importance of H-O actions, but several other actions that involve physical contact exist in our data that we have only started to explore. Some of them pertain to when two agents collaboratively manipulate objects. Yet others involve the HEL directly touching the ELD, as when helping the ELD with ambulating or with putting on clothing. More broadly, we need to better understand physical collaboration tasks and the role of communication in them.

Do humans interpret such communication at the symbolic level, or do we only use the signals to close the low-level control loops?

Finally, and equally important is to test the developed methodology on a robotic platform. A preliminary implementation of the dialogue act classifier (Ahmed 2014) has been completed in ROS (Quigley et al. 2009). We are in the process of implementing real-time variants of the H-O action recognition algorithms and experiments with a Nao robot (Gouaillier et al. 2009) are under development. An important question to address is also whether H-O actions can inform coreference resolution and dialogue act classification even when they are recognized automatically and thus with lower accuracy.

Acknowledgments

We thank former collaborators on RoboHelper: J. Ben-Arie, M. Foreman, L. Chen, S. Franzini, S. Jagadeesan, M. Javaid, and K. Ma. This work was originally supported by NSF award IIS-0905593. Additional funding was provided by NSF awards CNS-0910988, CNS-1035914 and IIS-1445751.

References

Ahmed, U. 2014. Implementing the Robot Modules in ROS. Master's thesis, University of Illinois at Chicago.

Anderson, A. H.; Bader, M.; Bard, E. G.; Boyle, E.; Doherty, G.; Garrod, S.; Isard, S.; Kowtko, J.; McAllister, J.; Miller, J.; Sotillo, C.; Thompson, H. S.; and Weinert, R. 1991. The HCRC Map Task Corpus. *Language and Speech* 34(4):351–366.

Barron-Gonzalez, H., and Prescott, T. 2014. Discrimination of Social Tactile Gestures Using Biomimetic Skin. In Natraj, A.; Cameron, S.; Melhuish, C.; and Witkowski, M., eds., *Towards Autonomous Robotic Systems*, Lecture Notes in Computer Science. Springer Berlin Heidelberg. 46–48.

Chen, L., and Di Eugenio, B. 2012. Co-reference via Pointing and Haptics in Multi-Modal Dialogues. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 523–527.

Chen, L., and Di Eugenio, B. 2013. Multimodality and Dialogue Act Classification in the RoboHelper Project. In *SIGDIAL*, 183–192.

Chen, L.; Javaid, M.; Di Eugenio, B.; and Žefran, M. 2015. The roles and recognition of haptic-ostensive actions in collaborative multimodal human–human dialogues. *Computer Speech & Language* 34(1):201231.

Chen, L. 2014. *Towards Modeling Collaborative Task Oriented Multimodal Human-human Dialogues*. Ph.D. Dissertation, University of Illinois at Chicago.

Cohen, P., and McGee, D. 2004. Tangible multimodal interfaces for safety-critical applications. *Communications of the ACM* 47(1):41–46.

Foster, M.; Bard, E.; Guhe, M.; Hill, R.; Oberlander, J.; and Knoll, A. 2008. The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, 295–302. ACM.

Gouaillier, D.; Hugel, V.; Blazevic, P.; Kilner, C.; Monceaux, J.; Lafourcade, P.; Marnier, B.; Serre, J.; and Maisonnier, B. 2009. Mechatronic design of NAO humanoid. In *ICRA '09, IEEE International Conference on Robotics and Automation*, 769–774.

Graff, D.; Canavan, A.; and Zipperlen, G. 1998. Switchboard-2 Phase I. LDC 99S79– http://www.ldc.upenn.edu/Catalog.

Jaimes, A., and Sebe, N. 2007. Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding* 108(1-2):116–134.

Javaid, M.; Žefran, M.; and Di Eugenio, B. 2014. Communication through physical interaction: A study of human collaborative manipulation of a planar object. In *RO-MAN 2014: The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 838–843.

Javaid, M. 2014. *Communication through Physical Interaction: Robot Assistants for the Elderly*. Ph.D. Dissertation, University of Illinois at Chicago.

Kipp, M. 2001. Anvil-a generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, 1367–1370.

Krapp, K. M. 2002. *The Gale Encyclopedia of Nursing & Allied Health*. Gale Group, Inc. Chapter Activities of Daily Living Evaluation.

Levin, L.; Thymé-Gobbel, A.; Lavie, A.; Ries, K.; and Zechner, K. 1998. A discourse coding scheme for conversational Spanish. In *Fifth International Conference on Spoken Language Processing*, 2335–2338.

Noda, T.; Ishiguro, H.; Miyashita, T.; and Hagita, N. 2007. Map acquisition and classification of haptic interaction using cross correlation between distributed tactile sensors on the whole body surface. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1099–1105.

Quigley, M.; Conley, K.; Gerkey, B.; Faust, J.; Foote, T.; Leibs, J.; Wheeler, R.; and Ng, A. Y. 2009. ROS: an open-source robot operating system. In *ICRA Workshop on Open Source Software*.

Sakoe, H., and Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing,* 26(1):43–49.

Shriberg, E.; Dhillon, R.; Bhagat, S.; Ang, J.; and Carvey, H. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proceedings of 5th SIGdial Workshop on Discourse and Dialogue*, 97–100.

Silvera-Tawil, D.; Rye, D.; and Velonaki, M. 2014. Interpretation of Social Touch on an Artificial Arm Covered with an EIT-based Sensitive Skin. *International Journal of Social Robotics* 6(4):489–505.