Joint Reference and Intention in Human-Robot Collaboration

Guy Hoffman and Andrea Lockerd

1. Introduction

In this project we would like to tackle the problems of joint reference and joint intention in a human-robot collaboration task. Our work is motivated by the desire to teach a robot (Leo) a task from beginning to end and then have him act in a robust way to execute this task in collaboration with a human teammate. In order to do so, both the problems of joint reference and of joint intention need to be solved.

Andrea will be working on the teaching stage of the task. In it, the human may be talking about a particular object, a feature of an object, or a whole group of objects. Leo will need to be able to attach labels and refer to all of these different levels of description in order to appropriately learn the task.

Guy will be working on the collaborative problem-solving stage. In it, we would like to demonstrate the principles of I-intentions as derivatives of We-intentions and the creation of subplans in the robotic agent based on a self-estimate of his capabilities. We would also like to show a dynamic adjustment of subplan meshing and sequencing according to failures and changing conditions in the world and the human side of the interaction.

The solution to these problems go hand-in-hand, as the robotic agent needs to have a correct representation of the task and its elements in order to form useful beliefs and goals related to that task.

2. Joint Reference

2.1 Problem Statement

Tackling the issue of 'aboutness', one part of our project will involve inferring and establishing joint reference at different scales using multiple modalities. This involves attributing the right reference level to social cues instead of hard-wiring attention to the object level or a particular salient object that's being referenced. A symbol (gesture or speech) can refer to an object, a group of objects, a task, a feature of an object, an action or a sequence of actions; a socially aware agent needs to be able to distinguish these reference frames in order to collaborate correctly with a human partner.

Practically, the problem can be framed in terms of Leo's implementation. The vision system is constantly providing a collection of possibly salient aspects of the visual environment, and the speech system has information about labels and commands. When something happens (i.e. Leo hears 'This is X', or sees someone perform an action), what is that label or that action 'about'; to what should he be 'paying attention'?

In the current implementation of Leo's behavior system, symbol attachment is fairly simplistic. Upon hearing 'This is X' the label 'X' is attached to the object that a point gesture is referencing. One goal in this project is to address this problem of symbol attachment, giving Leo a more flexible way of attaching labels to representations of things in the world.

2.2 Proposed Implementation

Baron-Cohen (1991) suggests that a joint attention mechanism works in the following way: make all things have a valence tag indicating whether or not 'I' think the other person finds this thing interesting or attention-worthy; continually monitor visual attention and actions to decide if something's attention valence has changed.

While fairly abstract, this theory suggests a three-part implementation of a joint reference mechanism. First is the ability to identify all of the possible references. Second is to assign some likelihood to each of these as being the actual reference at a given time. Third is the idea that the human partner can help determine accuracy. This final aspect, of monitoring the other person to see if you are correct, is the point that is left out of many cognitive models.

There are two cognitive models under consideration for this implementation, and both are explained briefly below. In relation to this project, the most interesting point is looking at how either of these cognitive models could be improved through this monitoring of the human partner.

One model is a Bayesian approach where the right level of reference is hypothesized to be one of many choices in a hypothesis space. And then the right level is learned through a few examples (Tenebaum, 2000). In this model, the human partner would somehow help to constrain the hypothesis space in consideration in order to speed up the learning process.

Kruschke (2003) has an interesting connectionist approach that he talks about as *attention in learning*. In this approach, weights are learned between all of the percepts and a label. What is learned through the examples is how all of the percepts (or *cues* as he calls them) contribute to or predict the label. This would work well for labeling features or groups of features like `button1, button2, on, off, red, green`. It is less clear how this would extend to labels that are groups of objects or sequences of actions.

The first big question in both of these approaches is how the hypothesis space should be formed. In this implementation all percepts will be primitive features at the lowest level of possible reference. From the vision and speech systems we are assuming the following feature space:

- how many people are in front of Leo: 1 2 or 3
- head pose or gaze direction of person, depending on the status of louis felip's module
- whether or not they are holding an object
- speech volume threshold
- nautilus speech parse
- for every 'object': location x,y,z, color, size, shape - rectangles, circles, squares

- buttons in particular have an on/off feature
- possible future speech percepts: direction of sound, prosidy of speech

As a first shot, all of the percepts and groups of percepts will be considered possible levels of attention. They fit into the following hierarchy:

1) features
2) groups of features
3) objects
4) groups of objects

This type of a hierarchy will work well for learning static references where a reference does not have any temporal or causal information. Static references can then be used in distinguishing between labeling an object, labeling a feature of an object, and labeling a state of the world.

Things get more complicated when the goal-state is dynamic (pushing buttons in a particular sequence). While we aspire to eventually have Leo reference a dynamic sequence of states and actions using the same mechanisms, we will treat the static and dynamic cases separately at first.

Once we have the ability to reference different levels of static states, we will need to add the following to our hierarchy of attention in order to reference dynamic events.

5) actions
6) sequences of actions
7) sequences of states of the world

2.3 Relevant Course Literature

One of the precursors to Theory of Mind (ToM), or reasoning about another person's mental states, is being able to hold joint attention; the ability to decide what a person is attending to or referring to. A number of people have suggested that joint attention is one of the first ToM capabilities seen in children (Butterworth, 1994)(Wellman, 1991). Understanding what is meant by a pointing gesture is to understand that the other person is focusing their attention on an object or an event. Baron-Cohen (1991) suggests then that understanding and reasoning about another person's attention is a precursor to understanding their beliefs.

3. Joint Intention

3.1 Problem Statement

In a collaborative task, a number of agents work together to solve a common problem. For this to take place, a joint course of action must emerge from the collection of the individual actions of the agents. In human collaboration, this does not reduce to just the sum of the individual actions derived from individual intentions, but rather is an interplay of actions inspired and affected by a joint or group intention.

Several models have been proposed to explain how joint intention and action relate to individual intentions and actions. The most salient factors crucial to the establishment of a robust collaborative framework seem to be: *communication, commitment to the joint task, commitment to mutual support, trust* and *dynamic meshing of subplans and action steps.* Particularly, it has been claimed that a robust collaboration scheme in a changing environment with partial knowledge and beliefs requires all of the above to allow for adequate execution and error-recovery.

3.2 Proposed Implementation

An important goal of our experiment is to empirically test some of the theoretical claims regarding joint intention (mutual support, commitment, communication etc.) We will hopefully demonstrate a human-robot collaborative task in which the robotic agent holds a set of individual intentions as a function of (a) the common goal of the team, (b) the agent's beliefs about the human collaborator's intentions and actions, (c) the agent's belief about the world state and (d) the agent's belief about his own capabilities.

It is important to note that we hold a goal-centric view as a fundamental feature of the ability to break down a common intention into individual intentions.This is true not only for the ability to view a joint action with respect to a particular goal, but also for the definition of individual intents based on sub-goals of the common intention.

To this end, Leo will create individual intentions based on both his understanding of the common goal of the team and his understanding of his own capabilities. He will be able to communicate with his human teammate about the commencement and completion of task steps.The robot will be able to recognize changes in the task environment as well as successes and failures on both his and his teammate's side. And most importantly, the robot will be able to communicate to his teammate the successful completion or unattainability of a crucial task step or the complete joint action. Moreover, the robot will be able to trust the human collaborator to communicate her own views regarding these very changes of state.

As a result of this study, we hope to get a clearer understanding of the role various elements (such as communication through expression, trust, commitment and mutual support) play in the creation and maintenance of individual intentions as part of joint intentions and common goals.

3.3 Relevant Course Literature

The models of joint intention that we will attempt to put to empirical test are based on the philosophical groundwork laid out in the early 1990s. Searle (1990) claims that collective intent and action cannot be formalized as a function of the individual intentions of the agents involved, but rather that the individual intentions are derived from their role in the common goal. He also stresses the importance of a social infrastructure to support collaboration. Bratman (1992) breaks down Shared Cooperative Activity into mutual responsiveness, commitment to the joint activity and commitment to mutual support. He also introduces the idea of meshing subplans, which our project will attempt to generalize to dynamically meshing subplans.

In the field of computer science, Cohen and Levesque (1990, 1991) establish the notions of trust and communication for robust collaborative activity and propose a formal framework to allow for the derivation of individual intent from a joint goal, effective error-recovery and dynamic adjustment of goals.

4. Demonstration and Criteria

We propose the following demonstration to illustrate and investigate the above points. Leo will interact with a human collaborator to learn and then solve a problem that is simple in terms of motor skills, but involves a number of sequenced steps as well as an understanding of various reference frames. Such a task could be sorting a number of blocks into two heaps according to color or (utilizing Leo's current motor abilities) playing Simon/entering a PIN by pressing buttons in a predetermined order. For the sake of the following description let's assume that the task is to turn all the buttons on and then to turn them all off.

4.1 Learning and Joint Reference

The following is a description of ways we will be able to demonstrate the new abilities of Leo's attention mechanism in the learning stage of this demonstration.

Leo should be able to interact with a person and have the following labels get attached to the object class level and to the particular object level:

```
Leo this is <a button>
Leo this is <button 1>
Leo this is <button 2>
```

The following questions should verify his having made the right attachments:

```
Is this one a button?
Where is button 2?
```

Alternatively, in an interaction the following labels should get attached to the feature level:

```
This button is <red>
Button 2 is <on>
This button is <off>
```

The following questions should verify that the right attachment has been made:

```
Is this button off?
Is this a red button?
Press button 1 until <it is on>
```

Then on a higher level Leo should be able to talk about states of groups of buttons.

```
Which buttons are on?
Which buttons are red?
```

Using these joint references and learning scenarios, Leo will build a representation of the task he and the human agent need to solve in collaboration. At this point, the joint attention and sub-planning model will come into play.

4.2 Joint Intention and Meshing Subplans

At every stage, either the human will do her part in the task or Leo will do his. While in a more general framework simultaneous action is appropriate, we will implement a turn-taking approach to avoid delving into the motor and perception challenges associated with concurrent action.

Leo will derive his own intentions based on the joint intention of the team, as well as based on his assessment of his own reaching ability. As the task progresses, Leo will be aware of the current state of completion both of the common goal and his own goals. This will be evident by the types of actions he will choose to do at any step. If the human chooses to take on one of Leo's tasks, Leo must adjust his plans to fit this new state of affairs. If Leo fails to accomplish a task, he will indicate this to his human collaborator "asking" her to help him. This can be done using a version of imperative pointing derived from Leo's already existing pointing gesture (which is, actually, a trimmed down version of his reaching gesture - just as early imperative pointing in infants is a reaching gesture to an object out of reach).

We will focus our attention on this and related communication abilities, with particular emphasis on the communication of success, failure and unattainability of the task as a whole and steps in this task. These will be tested by examples of the human agent "giving up on the task", her signaling the robot to do something he can't do, and bilateral communication of task completion.

A successful implementation will have Leo correctly and dynamically formulate and act on his goals as a function of the common goals, his reach and the human's temperamental changes of mind. He will be able to read his collaborator's turn-taking cues and adjustments as well as communicate his own beliefs about the current state of the world and his goals.

Looking at the implementation of this undertaking, this project will be roughly based on Leo's current motor capabilities, allowing Leo to point to buttons, look at buttons and the human teammate, and insert expressions involving any degree of freedom mechanically active. It will also build on Leo's current perception system, being able to read human gestures and the state of his button's on/off state. Since these are all in place, we will probably prefer demonstrating a button-pressing task rather than a block-sorting task, which will involve new work on both the motor and the perceptual front. This might still change depending on other students' proposal with respect to the motor and perceptual modules.

Leo's intention system will be a joint-intention model, which dynamically assigns tasks between the members of the collaboration team . Seeing everything through the common task filter will allow the robot to dynamically change his aspect of the task while still keeping in mind the common goals.

We plan to utilize the C5M belief system to model beliefs about the world, but will probably resort to a new belief representation model for beliefs about the human's state of mind and the common goal. The precise workings of these models are still to be defined.

5. Conclusions

In conclusion, we plan to implement joint reference and joint intention capabilities on Leo. This involves the ability to learn a task jointly with a person, attending to the right level of reference to build an accurate representation of the task. Given this task representation, we will demonstrate Leo's ability to collaborate with a partner to execute it. The current demonstration we are working towards involves Leo learning the task of turning all of his buttons on and off, then demonstrating the ability to perform this task jointly with a person.

6. References

S. Baron Cohen (1991), "Precursors to a Theory of Mind: Understanding Attention in Others ." In A. Whiten (ed) Natural Theories of Mind . Blackwell. Chapter 16.

M. Bratman (1992). " Shared Cooperative Activity ," The Philosophical Review , 101(2) pp. 327-341.

G. Butterworth (1994), " Theory of Mind and the Facts of Embodiment ," In C. Lewis and P. Mitchell (eds.), Children's Early Understanding of Mind. Lawrence Erlbaum Assoc. Chapter 6.
P. Cohen (1991). " Teamwork ," Nous 25 , pp 487-512.

J. K. Kruschke (2003), "Attention in Learning", In Current Directions in Psychological Science, 12, 171-175.

H. Levesque, P. Cohen, J. Nunes (1990), " On Acting Together ," In Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90) , pp. 94--99, Boston, MA, 1990.

J. Searle (1990). " Collective Intentions and Actions ," In Cohen, Morgan and Pollack (eds.) Intentions in communication . MIT Press, Chapter 19.

J. B. Tenenbaum, F. Xu (2000), "Word learning as Bayesian inference," In Proceedings of the 22nd Annual Conference of the Cognitive Science Society.

H. Wellman (1991), " From Desires to Beliefs: Acquisition of a Theory of Mind ." In A. Whiten (ed.) Natural Theories of Mind . Blackwell. Chapter 2.