Going Beyond Literal Command-Based Instructions: Extending Robotic Natural Language Interaction Capabilities

Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz

Human-Robot Interaction Laboratory
Tufts University, Medford, MA, USA
{thomas_e.williams, gordon.briggs, bradley.oosterveld, matthias.scheutz}@tufts.edu

Abstract

The ultimate goal of human natural language interaction is to communicate intentions. However, these intentions are often not directly derivable from the semantics of an utterance (e.g., when linguistic modulations are employed to convey politeness, respect, and social standing). Robotic architectures with simple command-based natural language capabilities are thus not equipped to handle more liberal, yet natural uses of linguistic communicative exchanges.

In this paper, we propose novel mechanisms for inferring intentions from utterances and generating clarification requests that will allow robots to cope with a much wider range of task-based natural language interactions. We demonstrate the potential of these inference algorithms for natural human-robot interactions by running them as part of an integrated cognitive robotic architecture on a mobile robot in a dialogue-based instruction task.

Introduction

When humans interact in natural language (NL) as part of joint activities, their ultimate goal is to understand each others' intentions, regardless of how such intentions are expressed. While it is sometimes possible to determine intentions directly from the semantics of an utterance, often the utterance alone does not convey the speaker's intention. Rather, it is only in conjunction with goal-based, task-based, and other context-based information that listeners are able to infer the intended meaning, such as in *indirect speech acts* where requests or instructions are not apparent from the syntactic form or literal semantics of the utterance. Given that an important goal of human-robot interaction is to allow for *natural interactions* (Scheutz et al. 2007), robotic architectures will ultimately have to face the challenge of coping with more liberal and thus natural human speech.

Enabling a broader coverage of human speech acts (beyond imperatives expressing commands), however, is quite involved and requires various additional mechanisms in the robotic architecture. In this paper, we introduce novel algorithms based on Dempster-Shafer (DS) theory (Shafer 1976) for inferring intentions I from utterances U in contexts C, and, conversely, for generating utterances U from intentions I in contexts C. We select more general DS-based representations over single-valued probabilities because the probability-based Bayesian inference problem to calculate $P(I|U,C)$ in terms of $P(U|I,C)$ is not practically feasible, for at least two reasons: (1) we do not have access to distributions over an agent's intentions (as we cannot look inside its head), and (2) we would need a table containing priors on all combinations of intentions and contexts. Instead, we employ rules of the form $u \wedge c \rightarrow_{[\alpha,\beta]} i$ that capture intentions behind utterances in particular contexts, where $[\alpha,\beta]$ is a confidence interval contained in [0,1] which can be specified for each rule independently (e.g., based on social conventions, or corpora statics when available). These rules are very versatile in that they can be defined for individual utterances and contexts or whole classes of utterances and contexts. Most importantly, we can employ DS-based modus ponens to make uncertain deductive and abductive inferences which cannot be made in a mere Bayesian framework. For more details justifying this approach, see (Williams et al. 2014).

We start with background information on instruction-based robotic architectures and basic Dempster-Shafer theoretic concepts, and then introduce the proposed algorithms for pragmatic inference and for generating requests to disambiguate intended meanings. Then we demonstrate the operation of the algorithms in a detailed example showing how uncertainty is propagated at each stage of processing and can lead to different responses by the robot. We finish with a brief discussion of the proposed approach and possible directions for future work.

Previous Work

Over the last several years, various integrated robotic architectures with natural language capabilities have been proposed for instruction-based natural language interactions (Lemaignan et al. 2014; Kruijff et al. 2010; Chai et al. 2014; Deits et al. 2013; Scheutz et al. 2013; Jing et al. 2012). Some of these approaches (e.g. (Jing et al. 2012)) are focused on compiling low-level controls from high-level, natural language commands and constraints and do not address efficient *real-time* interaction. Other architectures, while concerned with real-time interactions, do not generally perform pragmatic analyses to infer non-literal meanings from received utterances (and thus are not able to systematically handle utterances whose literal semantics do not directly reflect their speaker's intentions). For instance, while the architecture in (Deits et al. 2013) is able to use uncertainty to resolve references in received utterances, it does not do similar reasoning to resolve non-literal intentions. Likewise, (Chai et al. 2014) is focused on reference resolution in the light of referential uncertainty but not on non-literal intention understanding. Pragmatic inference *is* performed by the integrated architecture presented in (Scheutz et al. 2013), but that architecture does not explicitly represent the uncertainty of the robot's knowledge, and thus their pragmatic inference components are not robust to uncertain context, input, or pragmatic rules. One integrated approach does handle pragmatic inference with an explicit representation of uncertainty (Wilske and Kruijff 2006), but is limited to handling indirect commands and uses a rather rudimentary representation of uncertainty. The goal of this paper is to tackle currently unaddressed challenges posed by more liberal human language usage. This requires not only the addition of several components for handling aspects of natural language pragmatics, but also representational and inferential mechanisms to robustly capture and handle the uncertainties that plague natural, real-world communication.

Basic Notions of Dempster-Shafer Theory

Since the proposed alogorithms and architecture will use DS-theoretic representations of uncertainty, we briefly review the basic concepts of this framework for reasoning about uncertainty, which is a generalization or extension of the Bayesian framework (Shafer 1976).

Frame of Discernment: A set of elementary events of interest is called a *Frame of Discernment* (FoD). A FoD is a finite set of mutually exclusive events $\Theta = \theta_1, ..., \theta_N$. The power set of Θ is denoted by $2^\Theta = A : A \subseteq \Theta$.

Basic Belief Assignment: Each set $A \subseteq 2^\Theta$ has a certain weight, or *mass* associated with it. A *Basic Belief Assignment* (BBA) is a mapping $m_\Theta(\cdot) : 2^\Theta \to [0, 1]$ such that $\sum_{A \subseteq \Theta} m_\Theta(A) = 1$ and $m_\Theta(\emptyset) = 0$. The BBA measures the support assigned to the propositions $A \subseteq \Theta$ only. The subsets of A with non-zero mass are referred to as *focal elements* and comprise the set F_Θ. The triple $E = \{\Theta, F_\Theta, m_\Theta(\cdot)\}$ is called the *Body of Evidence* (BoE).

Belief, Plausibility, and Uncertainty: Given a BoE $\{\Theta, F_\Theta, m_\Theta(\cdot)\}$, the *belief* for a set of hypotheses A is $Bel(A) = \sum_{B \subseteq A} m_\Theta(B)$. This belief function captures the total support that can be committed to A without also committing it to the complement A^c of A. The *plausibility* of A is $Pl(A) = 1 - Bel(A^c)$. Thus, $Pl(A)$ corresponds to the total belief that does not contradict A. The *uncertainty* interval of A is $[Bel(A), Pl(A)]$, which contains the true probability $P(A)$. In the limit case with no uncertainty, we get $Pl(A) = Bel(A) = P(A)$.

Inference and Fusion: Uncertain logical inference can be performed using DS-theoretic modus ponens (denoted $\odot$) (Tang et al. 2012). We will use the DS-theoretic AND (denoted $\otimes$) to combine BoEs on different FoDs (Tang et al. 2012), and Yager's rule of combination (denoted $\bigcap$) to combine BoEs on the same FoD (Yager 1987). We choose to use Tang's models of modus ponens and AND over other proposed models due to the counter-intuitive results of those models, and because those models do not allow uncertainty to be multiplicatively combined. More details about the strengths and weaknesses of various options for DS-theoretic uncertain inference can be found in (Williams et al. 2014). Yager's rule of combination is chosen because it allows uncertainty to be pooled in the universal set, and due to the counter-intuitive results produced by Dempster's rule of combination (as discussed in (Zadeh 1979)).

Logical AND: For two logical formulae ϕ_1 (with $Bel(\phi_1) = \alpha_1$ and $Pl(\phi_1) = \beta_1$) and ϕ_2 (with $Bel(\phi_2) = \alpha_2$ and $Pl(\phi_2) = \beta_2$, applying logical AND yields $\phi_1 \otimes \phi_2 = \phi_3$ with $Bel(\phi_3) = \alpha_1 * \alpha_2$ and $Pl(\phi_3) = \beta_1 * \beta_2$.

Modus Ponens: For logical formulae ϕ_1 (with $\mathrm{Bel}(\phi_1) = \alpha_1$ and $Pl(\phi_1) = \beta_1$) and $\phi_{\phi_1 \to \phi_2}$ (with $Bel(\phi_{\phi_1 \to \phi_2}) = \alpha_R$ and $Pl(\phi_{\phi_1 \to \phi_2}) = \beta_R$, the corresponding model of modus ponens is $\phi_1 \odot \phi_{\phi_1 \to \phi_2} = \phi_2$ with $Bel(\phi_2) = \alpha_1 * \alpha_R$ and $Pl(\phi_2) = Pl(\phi_R)$.

Measuring Uncertainty: We will use the "uncertainty measure" λ discussed in (Williams et al. 2014) to compare the uncertainties associated with formulae ϕ and their respective confidence intervals $[\alpha, \beta]$:

$$\lambda(\alpha, \beta) = 1 + \frac{\beta}{\gamma} log_2 \frac{\beta}{\gamma} + \frac{1 - \alpha}{\gamma} log_2 \frac{1 - \alpha}{\gamma}$$

$$\text{where } \gamma = 1 + \beta - \alpha.$$

Here, ϕ is deemed more uncertain as $\lambda(\alpha, \beta) \to 0$. We introduce an "uncertainty threshold" Λ (set to 0.1) where utterances with $\lambda(\alpha, \beta) < \Lambda$ will require clarification from an interlocutor.

For more details on our use of Dempster-Shafer theoretic uncertain logical inference, we direct readers to (Williams et al. 2014)

Algorithms and Architecture

A cognitive robotic architecture capable of going beyond direct command-based instructions needs several high-level components in addition to typical NL components (such as speech recognizers, parsers, etc.) that work in concert to extract intended meanings and generate informative clarification questions and feedback as needed. Figure 1 depicts how these new components are inserted into the architecture we extend (i.e., the Distributed Integrated Affect, Recognition and Cognition (DIARC) architecture (Scheutz et al. 2007)).

When an interlocutor speaks to the robot, speech is processed via the standard NL pipeline (speech recognizer, syntactic and semantics parser) resulting in candidate semantic expressions ϕ, each with its own uncertainty interval $[\alpha, \beta]$ attached. While a typical command-based system (e.g., (Dzifcak et al. 2009)) would attempt to act on the semantic interpretation with the highest confidence (and fail if it is not actionable), in the proposed architecture semantic representations are further processed in a pragmatic inference component, which attempts to apply modulatory pragmatic rules

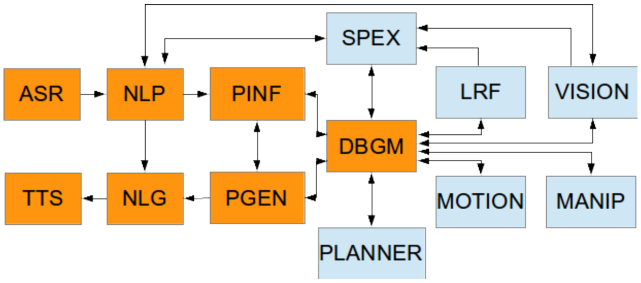

Figure 1: partial architecture diagram. Highlighted are the components that form the natural language pipeline: Automatic Speech Recognition (ASR), Natural Language Processing (NLP), Natural Language Generation (NLG), Text-to-Speech (TTS), Pragmatic Inference (PINF), Pragmatic Generation (PGEN), and Dialogue, Belief and Goal Management (DBGM). Also shown are relevant components that interact with the DBGM: the SPatial EXpert (SPEX), Task Planner (PLANNER), Motion Planner (MOTION), Manipulation (MANIP), Laser Range Finder (LRF), and Vision (VISION).

to utterance and semantic representations to infer the intentions of the speaker.

The semantic interpretation is passed to our new component for Pragmatic Inference (PINF), which uses contextual and general knowledge to determine the *intention* underlying the literal semantics. By using pragmatic rules indexed by utterance and context, PINF can determine, for example, that asking if one knows the time should be interpreted not as a "yes-no question", but as an indication that the speaker would like to be told what time it is. The resulting intention or intentions can then be returned to DIARC's "Dialogue, Belief, and Goal Manager" (DBGM), which is responsible for dialogue management (Briggs and Scheutz 2012), storing beliefs in its knowledge base, performing inference on those beliefs, tracking and managing goals (Brick, Schermerhorn, and Scheutz 2007; Scheutz and Schermerhorn 2009), and determining what actions to take in pursuit of its goals. If an utterance is determined to be a command or request, the DBGM will instantiate a new goal and determine how best to accomplish that goal. If the DBGM determines that it should respond to an interlocutor, the intention it desires to communicate is passed to our new component for Pragmatic Generation (PGEN), which determines the best way to effectively and politely communicate that intention. From this point on, information flows in the standard fashion through natural language generation and speech synthesis. Next, we will provide the details for the core algorithms of PINF and PGEN.

Pragmatic Inference

The goal of pragmatic analysis is to infer intentions based on (1) the semantics of incoming utterances, (2) the robot's current context, and (3) the robot's general knowledge. Our pragmatic inference algorithm (originally presented in (Williams et al. 2014), is shown in Algorithm 1. This algorithm takes three parameters: (1) a BoE of candidate utterances $\{\Theta_U, m_u\}$ provided by NLP, (2) a BoE of relevant contextual items $\{\Theta_C, m_c\}$ provided by the DBGM, and (3) a table of pragmatic rules R. Each rule $r_{uc \to i}$ in R is indexed by an utterance u and a set of contextual items c, and dictates the mass assigned to $Bel(i)$ and $Pl(i)$ when the robot believes that utterance u was heard and that contextual items c are true. Here i is a logical formula representing the intention the interlocutor was expressing through utterance u. When these contextual items involve the shared context of the robot and its interlocutor, they are couched in terms of the *interlocutor's beliefs*. This is critical, as the intentions of the robot's interlocutor are dependent not on the robot's beliefs, but on his or her own beliefs. This allows the robot to correctly interpret its interlocutor's intentions when cognizant of discrepancies between its own beliefs and its interlocutor's beliefs, and to identify information of which its interlocutor may want to be informed. This is important for both pragmatic inference and generation, as this paradigm implicitly assumes that the robot's interlocutor communicates according to the same table of rules known to the robot (however, it is straightforward to keep separate rule tables for individual interlocutors if required).

Algorithm 1 getIntendedMeaning($\{\Theta_U, m_u\}$,$\{\Theta_C, m_c\}$,R)

```
1:  $\{\Theta_U, m_u\}$: BoE of candidate utterances
2:  $\{\Theta_C, m_c\}$: BoE of relevant contextual items
3:  $R$: Currently applicable rules
4:  $S = \emptyset$
5:  for all $r \in R$ do
6:      $S = S \cup \{(m_u \otimes m_c) \odot m_{r=uc \to i}\}$
7:  end for
8:  $G = group(S)$
9:  $\psi = \emptyset$
10: for all group $g_i \in G$ do
11:     $\psi = \psi \cup \{\bigcap_{j=0}^{|g_i|} g_{i_j}\}$
12: end for
13: return $\psi$
```

When an utterance is heard, each rule $r_{uc \to i} \in R$ is examined (line 5), and m_{uc} is determined by performing $m_u \otimes m_c$ (line 6), where m_u specifies the degree to which utterance u is believed to be heard, and m_c specifies the degree to which each of the rule's associated contextual items is believed to be true. DS-based modus ponens is then used to obtain m_i from $m_{uc \to i}$ and m_{uc} (line 6).

While previous approaches (e.g., (Briggs and Scheutz 2013)) look for a single applicable rule in order to produce a single likely intention, we instead consider all applicable rules. This is particularly important for complex contexts or abstract context specifications, where multiple rules might be applicable. Moreover, the robot might have rules that apply to its particular context as well as to a more general context, and it may be more appropriate to consider the combined implicatures of all applicable rules rather than only considering, for example, the most specific applicable rule.

Since we may consider multiple rules, multiple intentions may be produced. And multiple rules may also produce the same intentions, possibly with different levels of belief or disbelief. To be able to generate a set of *unique* intentions implied by utterance u after considering all applicable pragmatic rules, we thus group intentions that have the same semantic content but different mass assignments (line 8) and use Yager's rule of combination (line 11) to fuse each group of identical intentions, adding the resulting fused intention to set ψ. This set then represents the set of intentions implied by utterance u and is returned to the DBGM, where its level of uncertainty is assessed: for each intention $i \in \psi$ on uncertainty interval $[\alpha_i, \beta_i]$, a clarification request is generated if $\lambda(\alpha_i, \beta_i) < \Lambda$. For example, consider a scenario in which the robot is unsure which of two contexts it is in. In the first context, a particular statement should be interpreted as a request for information, and in the second context, it should be interpreted as an instruction. In this case, the robot will ask "Should I <perform the intended action> or would you like to know <the intended information>?" This demonstrates the ability for the robot to exploit propagated uncertainty to identify and resolve uncertainties and ambiguities.

A similar process is also seen directly *before* pragmatic inference: after NLP produces set of surface semantics s, those semantic interpretations are analyzed using the λ ambiguity measure. If $\lambda(\alpha_P, \beta_P) < \Lambda$ for semantic predicate p with uncertainty interval $[\alpha_P, \beta_P]$, a request to verify what was said is sent to NLG, which generates and communicates a realization of the form "Did you say that <s>" in which case the uncertain semantics are *not* passed on for pragmatic analysis.

Pragmatic Generation

When the robot needs to communicate information, it must choose appropriate surface realizations of the semantic facts in intends to convey. However, for reasons of social convention such as politeness, it may be inappropriate to express semantic facts in the most direct manner. For example, one may find it rude if the robot were to say "I want to know what time it is. Tell me now." To allow the robot to generate socially acceptable utterances based on pragmatic considerations, we introduce an abductive inference algorithm called *pragmatic generation*, which, much like pragmatic inference, uses the robot's current context and its set of pragmatic rules to determine the best utterance to communicate intentions. The "best" utterance is determined to be the utterance that, according to the robot's set of pragmatic rules, would be most likely to communicate the given intention properly (e.g., without communicating any other information that the robot does not believe to be true). A DS-based approach is particularly useful here, because rule-based pragmatic inferences are determined by equations that relate the premise and rule to the consequent and can thus, exactly because they are equations, be used for inferences in both directions, deductive and abductive. We can thus infer the best utterance to convey a given intention in a given context from the same rules we use for inferring the best intention given an utterance in the same context. Moreover, we can perform pragmatic generation recursively: if a pragmatic rule matches the high-level structure of an utterance, it may be necessary to further abduce the best way to phrase individual clauses of the utterance that were left open by the high-level rule.

Algorithm 2 getSemantics($\{\Theta_I, m_i\}, \{\Theta_C, m_c\}, R$)

```
 1: $\{\Theta_i, m_i\}$: BoE of candidate intentions
 2: $\{\Theta_C, m_c\}$: BoE of relevant contextual items
 3: $R$: Currently applicable rules
 4: $S = \emptyset$
 5: for all $r \in R$ do
 6:     $u = (m_i \otimes m_c) \odot m_{r=uc \to i}$
 7:     for all $(b_s, b_v) \in getBindings(u)$ do
 8:         if $marked(b_v)$ then
 9:             $u = adapt(u, getSemantics($
                $buildBoE(b_s), \{\Theta_C, m_c\}, R))$
10:         end if
11:     end for
12:     $u' = checkEffects($
        $getIntendedMeaning(\{\Theta_U, m_u\}, \{\Theta_C, m_c\}, R))$
13:     $S = S \cup u'$
14: end for
15: return $S$
```

As with pragmatic inference, the pragmatic generation algorithm (see Algorithm 2) takes the robot's current context $\{\Theta_C, m_c\}$ and the set of currently applicable rules R. Instead of the BoE of possible incoming utterances $\{\Theta_U, m_u\}$, the algorithm takes a BoE of possible intentions desired to be communicated $\{\Theta_I, m_i\}$, as determined by the DBGM. For each applicable rule $r_{uc \to i} \in R$, the algorithm performs an uncertain modus ponens operation producing a BoE indicating which utterance would most likely generate the desired intention according to rule r (line 6).

The algorithm then examines the structure of the resulting utterance (line 7) to determine whether it should recurse on subsections of the utterance, recursing on the semantics b_s associated with each variable b_v marked as suitable for recursion. For example, for the utterance $Want(self, Know(self, or(X, Y)))$, it may be necessary to recurse on the portions of the utterance bound to X and Y. Once the results of any such recursions (line 9) are integrated into the representation of the utterance to communicate u, the set of intentions ψ that would be implied by utterance u are calculated (on line 12) by calling $getIntendedMeaning(\{\Theta_U, m_u\}, \{\Theta_C, m_c\}, R)$ (i.e., Algorithm 1) with the candidate utterance and the current context and rule set. The belief and plausibility of u are then modulated by $Bel(p_i)$ and $Pl(p_i)$ for $p_i \in \psi$. This prevents the robot from accidentally communicating some proposition that it does not actually believe to be true. Finally, the set of candidate utterances S is returned, from which an utterance is chosen to communicate, e.g., by choosing the candidate with the highest degree of belief.

Demonstration

To demonstrate the operation of the proposed inference algorithms for natural human-robot interactions, we consider a dialogue interaction that occurs as part of a Search-and-Rescue Task. The interaction starts with an interlocutor ("Jim") telling the robot "Commander Z needs a medical kit." The utterance and semantic representation produced by NLP for this statement is

$$Statement(Jim, self, needs(commander\_z, medkit))[\alpha, \beta].$$

We will now examine how the dialogue between Jim and the robot plays out under three different combinations of values for α and β, corresponding with low, medium, and high of uncertainty accrued by the early NL components (up to semantic parsing). These three conditions are denoted[1]

U_{low} (with uncertainty interval $[0.95, 1.00]$),
U_{med} (with uncertainty interval $[0.62, 0.96]$), and
U_{high} (with uncertainty interval $[0.31, 0.81]$).

Furthermore, we will assume three settings that differ with respect to the robot's assumptions regarding its interlocutor's beliefs about who is subordinate to whom. In the first case (denoted C_{jim}), the robot believes that Jim believes that the robot is subordinate to him. In the second case (denoted C_{robot}), the robot believes that Jim believes that he is subordinate to the robot. In the third case (denoted C_{unk}), the robot is unsure who Jim believes to be the subordinate between the pair of them. The differences in these scenarios are reflected in differences in the knowledge base of the robot at the start of the task:

C_{jim}	$locationof(breakroom, medkit)[0.80, 0.90]$
	$Believes(Jim, subordinate(self, Jim))[0.80, 0.90]$
	$Believes(Jim, subordinate(Jim, self))[0.10, 0.20]$
C_{robot}	$locationof(breakroom, medkit)[0.80, 0.90]$
	$Believes(Jim, subordinate(self, Jim))[0.10, 0.20]$
	$Believes(Jim, subordinate(Jim, self))[0.80, 0.90]$
C_{unk}	$locationof(breakroom, medkit)[0.80, 0.90]$
	$Believes(Jim, subordinate(self, Jim))[0.50, 0.60]$
	$Believes(Jim, subordinate(Jim, self))[0.40, 0.50]$

In all conditions, the robot uses the following set of pragmatic rules (here intentions are represented as "Goal" and intentions to know are presented as "ITK(A,B)", e.g., (Perrault and Allen 1980)):

1. $Stmt(A, B, Want(A, bring(C, D, E))) \rightarrow$
   $Goal(C, bring(C, D, E))[0.95, 0.95]$
2. $AskWH(A, B, or(C', D')) \rightarrow$
   $ITK(A, or(C', D'))[0.95, 0.95]$
3. $Stmt(A, B, Want(A, Know(A, C))) \rightarrow$
   $ITK(A, C)[0.85, 0.85]$
4. $Instruct(A, B, C) \rightarrow$
   $Goal(B, C)[0.90, 0.90]$
5. if $Bel(A, subordinate(B, A))$:
   $Stmt(A, B, needs(C, D)) \rightarrow$
   $Goal(B, bring(B, D, C))[0.80, 0.90]$
6. if $Bel(A, subordinate(B, A))$:
   $Stmt(A, B, needs(C, D)) \rightarrow$
   $not(ITK(A, locationof(E, D)))[0.80, 0.90]$
7. if $Bel(A, subordinate(A, B))$:
   $Stmt(A, B, needs(C, D)) \rightarrow$
   $ITK(A, locationof(E, D))[0.80, 1.00]$
8. if $Bel(A, subordinate(A, B))$:
   $Stmt(A, B, needs(C, D)) \rightarrow$
   $not(Goal(B, bring(B, D, C)))[0.80, 1.00]$

In U_{high}, $\lambda(0.31, 0.81) < 0.1$, so the robot responds "Did you say that Commander Z needs a medkit?" In U_{med}, $\lambda(0.62, 0.96) > 0.1$, and in U_{low} $\lambda(0.95, 1.00) > 0.1$, and thus the semantics are passed on to PINF, which yields the following intentions for each combination of U and C conditions:

	C_{jim}
U_{low}	$Goal(self, bring(self, medkit, commander\_z))[0.88, 0.95]$ $ITK(Jim, locationof(X, medkit))[0.05, 0.12]$
U_{med}	$Goal(self, bring(self, medkit, commander\_z))[0.88, 0.93]$ $ITK(Jim, locationof(X, medkit))[0.07, 0.12]$
	C_{robot}
U_{low}	$Goal(self, bring(self, medkit, commander\_z))[0.05, 0.12]$ $ITK(Jim, locationof(X, medkit))[0.88, 0.95]$
U_{med}	$Goal(self, bring(self, medkit, commander\_z))[0.07, 0.12]$ $ITK(Jim, locationof(X, medkit))[0.88, 0.93]$
	C_{unk}
U_{low}	$Goal(self, bring(self, medkit, commander\_z))[0.47, 0.67]$ $ITK(Jim, locationof(X, medkit))[0.33, 0.54]$
U_{med}	$Goal(self, bring(self, medkit, commander\_z))[0.50, 0.62]$ $ITK(Jim, locationof(X, medkit))[0.38, 0.50]$

These intentions are then passed to the DBGM, which performs different operations based on the uncertainty condition. In C_{unk}, the high level of uncertainty necessitates a clarification request, so the DBGM forms intention i:

$$ITK(self, or(ITK(Jim, locationof(X, medkit)),$$
$$Goal(self, bring(self, medkit, commander\_z))))[1.0, 1.0].$$

$Bel(i)$ and $Pl(i)$ are both 1.0, since the robot can be sure of its own intentions. Given i, PGEN produces:

$$ITK(self, or(Want(Jim, Know(Jim, locationof(X, medkit))),$$
$$Want(Y, bring(self, medkit, commander\_z))))[0.95, 1.0].$$

NLG then translates this intention to "Would you like to know where to find a medkit? or would you like me to bring commander z a medkit?"

Suppose Jim responds "I'd like to know where to find one." In U_{high}, $\lambda(0.31, 0.81) < 0.1$, so the robot responds "Did you say that you would like to know where a medkit is located?" Otherwise, PINF produces:

U_{low}	$ITK(Jim, locationof(X, medkit))[0.81, 1.0]$
U_{med}	$ITK(Jim, locationof(X, medkit))[0.52, 1.0]$

In U_{med}, no uncertainty is initially detected, but the *intention* of the utterance resulting from PINF is deemed too uncertain since $\lambda(0.52, 1.00) < 0.1$, so the robot asks for clarification: "Would you like to know where to find a medkit?" In U_{low}, this intention is not deemed uncertain since $\lambda(0.81, 1.00) > 0.1$, so the intention is instead added

[1] All beliefs and plausibilities listed in this section are rounded to two decimal places for the reader's convenience.

to the robot's set of beliefs. This behavior, and the actions that follow, are identical to how the robot responds to the original utterance in scenario C_{robot}. Since Jim has not yet been provided an answer to his question, the robot attempts to answer him. The robot first queries its knowledge base to determine if it knows the answer. If it had not known the location of a medkit, it would have generated a response with the semantics

Stmt(self,Jim,not(Know(self,locationof(X,medkit))))[1.0,1.0].

In this scenario, the robot does know the answer as it has $locationof(breakroom, medkit)[0.80, 0.90]$ in its knowledge base, so it forms an utterance with semantics

Stmt(self,Jim, locationof(breakroom,medkit))[0.8,0.9].

NLG then translates this to "A medkit is located in the breakroom."

Suppose the robot's interlocutor instead responded to the initial clarification request by saying "Bring him one." In U_{high}, the robot would respond by saying "Did you say that I should bring commander Z a medkit?" Otherwise, PINF produces:

$$\begin{array}{ll} U_{low} & Goal(self, bring(self, medkit, commander\_z))[0.86, 1.0] \\ U_{med} & Goal(self, bring(self, medkit, commander\_z))[0.55, 1.0] \end{array}$$

This intention is not deemed uncertain in either condition[2] so the intention is instead added to the robot's set of beliefs. This behavior, and the actions that follow, are identical to how the robot responds to the original utterance in scenario C_{jim}. The DBGM then determines which action will accomplish the goal and executes that action, setting forth to retrieve the medkit. A video of this interaction in operation on a Willow Garage PR2 robot can be viewed at *https://vimeo.com/106203678.*

Discussion and Future Work

The goal of the demonstration example on a real robot in a real-world setting was two-fold. First, we intended to show the potential of the proposed algorithms for making sound deductive and abductive pragmatic inferences based on human utterances and context that go beyond the direct interpretation of command-based instructions. And second, we wanted to demonstrate that the algorithms have been fully integrated into an operational cognitive robotic architecture (even though space limitations did not permit us to present any details on the DIARC architecture outside of the proposed algorithms). Yet, the demonstration is clearly not an evaluation and should not be taken as such. While an evaluation of the integrated system will eventually be critical, we believe that it would be premature at present given that we do not even know how to best evaluate such integrated systems (e.g., how many dialogue-based scenarios would we have to set up and how many pragmatic rules would we have to examine to be able to make a case about how well the system works and how could we be sure that the employed data was sufficient?). Instead, the current system can be seen as a proof-of-concept that the proposed algorithms do not only work in principle and isolation, but in real-time as part of an integrated robotic architecture.

As a next step towards a full evaluation in the future, we are interested in improving several aspects of the current system, including how pragmatic rules can be acquired in a way that does not require the system to learn from large data sets offline. Specifically, we are interested in using NL instructions to learn rules quickly and to use reinforcement methods (based on feedback from human interlocutors) to adapt the uncertainty values associated with the learned rules. This way of allowing for online learning of pragmatic interpretations will enable adaptive trainable systems that can quickly acquire new knowledge on the fly as is often required in human-robot interaction domains.

We are also interested in extending PINF with plan reasoning capabilities so that it can better interpret non-idiomatic indirect speech acts, and extending PGEN so that it can use Grice's conversational maxims when choosing which utterance to communicate (e.g., analogous to (Briggs and Scheutz 2013)).

Conclusion

In this paper, we presented algorithms for inferring and communicationg intentions, and for generating clarification requests. We described how the integration of the proposed algorithms into the a robot cognitive architecture affords capabilities and robustness in handling natural language interactions that go beyond command-based instructions. We also highlighted the benefits of using Dempster-Shafer theory by stepping through the performance of the proposed algorithms in a demonstration task, where the algorithms were run in real-time as part of the DIARC integrated robotic architecture on a PR2 robot. Future work will extend the robustness and scope of the algorithms and investigate different methods for learning pragmatic rules effectively.

Acknowledgments

This research was funded in part by ONR grants #N00014-11-1-0493, #N00014-14-1-0751, and #N00014-14-1-0144.

References

Brick, T.; Schermerhorn, P.; and Scheutz, M. 2007. 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*, 1423–1428.

Briggs, G., and Scheutz, M. 2012. Multi-modal belief updates in multi-robot human-robot dialogue interaction. In *Proceedings of 2012 Symposium on Linguistic and Cognitive Approaches to Dialogue Agents*.

Briggs, G., and Scheutz, M. 2013. A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*.

[2] One could argue that the uncertainty in U_{med} is high enough to warrant a clarification request. One may raise Λ to achieve such behavior, if so desired.

Chai, J. Y.; She, L.; Fang, R.; Ottarson, S.; Littley, C.; Liu, C.; and Hanson, K. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, HRI '14, 33–40. New York, NY, USA: ACM.

Deits, R.; Tellex, S.; Kollar, T.; and Roy, N. 2013. Clarifying commands with information-theoretic human-robot dialog. *JHRI*.

Dzifcak, J.; Scheutz, M.; Baral, C.; and Schermerhorn, P. 2009. What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Proceedings of the 2009 International Conference on Robotics and Automation*.

Jing, G.; Finucane, C.; Raman, V.; and Kress-Gazit, H. 2012. Correct high-level robot control from structured english. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, 3543–3544. IEEE.

Kruijff, G.-J. M.; Lison, P.; Benjamin, T.; Jacobsson, H.; Zender, H.; Kruijff-Korbayová, I.; and Hawes, N. 2010. Situated dialogue processing for human-robot interaction. In *Cognitive Systems*. Springer Berlin Heidelberg. 311–364.

Lemaignan, S.; Warnier, M.; Sisbot, A. E.; and Alami, R. 2014. Human-robot interaction: Tackling the ai challenges. Technical report.

Perrault, C. R., and Allen, J. F. 1980. A plan-based analysis of indirect speech acts. *Computational Linguistics* 6(3-4):167–182.

Scheutz, M., and Schermerhorn, P. 2009. Affective goal and task selection for social robots. In Vallverdú, J., and Casacuberta, D., eds., *The Handbook of Research on Synthetic Emotions and Sociable Robotics*. IGI Global. 74–87.

Scheutz, M.; Schermerhorn, P.; Kramer, J.; and Anderson, D. 2007. First steps toward natural human-like HRI. *Autonomous Robots* 22(4):411–423.

Scheutz, M.; Briggs, G.; Cantrell, R.; Krause, E.; Williams, T.; and Veale, R. 2013. Novel mechanisms for natural human-robot interactions in the diarc architecture. In *Proceedings of AAAI Workshop on Intelligent Robotic Systems*.

Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton University Press.

Tang, Y.; Hang, C.-W.; Parsons, S.; and Singh, M. P. 2012. Towards argumentation with symbolic dempster-shafer evidence. In *COMMA*, 462–469.

Williams, T.; Núñez, R. C.; Briggs, G.; Scheutz, M.; Premaratne, K.; and Murthi, M. N. 2014. A dempster-shafer theoretic approach to understanding indirect speech acts. *Advances in Artificial Intelligence*.

Wilske, S., and Kruijff, G.-J. 2006. Service robots dealing with indirect speech acts. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 4698–4703. IEEE.

Yager, R. R. 1987. On the dempster-shafer framework and new combination rules. *Information sciences* 41(2):93–137.

Zadeh, L. A. 1979. *On the validity of Dempster's rule of combination of evidence*. Electronics Research Laboratory, University of California.