Resolution of Referential Ambiguity in Human-Robot Dialogue Using Dempster-Shafer Theoretic Pragmatics

Tom Williams and Matthias Scheutz
Human-Robot Interaction Laboratory
Tufts University, Medford, MA, USA
williams@cs.tufts.edu, matthias.scheutz@tufts.edu

***Abstract*—Robots designed to interact with humans in realistic environments must be able to handle uncertainty with respect to the identities and properties of the people, places, and things found in their environments. When humans refer to these entities using *under-specified* language, robots must often generate *clarification requests* to determine which entities were meant. In this paper, we present recommendations for designers of robots needing to generate such requests, and show how a Dempster-Shafer theoretic pragmatic reasoning component capable of generating requests to clarify *pragmatic* uncertainty can also generate requests to resolve *referential* uncertainty when integrated with a probabilistic reference resolution component.**

I. Introduction

Imagine a robot named Cindy and a human named Bob. Cindy and Bob are working together in a disaster relief scenario, and have just left a kitchen containing two medical kits: one on a table, and one on a counter. After driving for a few minutes, Bob turns to Cindy and asks "Can you go back to the kitchen and grab the medical kit?"

To successfully fulfill Bob's request, Cindy must resolve two types of ambiguity. Bob's request is *pragmatically ambiguous* as it could be interpreted *directly* (as a literal question as to Cindy's abilities) or *indirectly* (as a command to Cindy). Bob's request is *referentially ambiguous* because it could refer to either the medical kit on the table or the one on the counter.

In previous work, we showed how Dempster-Shafer (DS)-theoretic pragmatic reasoning could be used to both identify sources of pragmatic ambiguity and generate pragmatically appropriate clarification requests [27] to resolve such ambiguity. However, that work could not resolve *referential* ambiguity, and assumed that information about all referents was stored in a single, centrally located knowledge base (c.f. [25]).

In this work, we demonstrate the integration of a DS-theoretic pragmatic reasoning component with a probabilistic reference resolution algorithm, and show how this integration allows a robot to identify, and generate clarification requests to resolve, *referential* ambiguity as well. This approach is uniquely tailored to human-robot interaction (HRI) contexts, as it produces human-preferred clarification requests that conform with the pragmatics of human-robot dialogue.

The remainder of this paper proceeds as follows: First, we discuss previous work on clarification request generation in HRI contexts. Next, we present the results of a human-subjects experiment in which previous findings regarding human preferences with respect to robot clarification request formulation are replicated and refined. Then, we present and evaluate an approach to clarification request generation designed to align with human preferences. Finally, we discuss possible directions for future work.

II. Background

In this section, we first discuss previous work on natural language generation and clarification request generation. We then critique that work in order to generate a set of hypotheses regarding human preferences that should be accounted for when designing language-capable robots.

A. Previous Work

There has been much previous work in developing general natural language generation (NLG) systems. For example, Reiter et al. present an NLG framework comprised of six stages: content determination, document structuring, aggregation, lexical choice, referring expression generation (REG), and realization [16]. It is unclear, however, whether such frameworks are well suited to *situated contexts* in which an agent is embedded in a complex, dynamic, environment rife with uncertainty and ambiguity [12]. In HRI, for example, NLG is often performed to *solicit* information, whereas in non-situated contexts it is more typically performed to *provide* information. In previous work, we thus proposed an *HRI-oriented* clarification request generation framework comprised of five stages: (1) uncertainty identification, (2) decision to communicate, (3) utterance choice, (4) surface realization, and (5) speech synthesis [26]. In this paper, we present an integrated approach that implements all five stages.

Clarification request generation itself has also attracted a large amount of research overall [15, 23], but relatively little in *situated* contexts such as human-robot interaction. Recently, some researchers have used information-theoretic techniques to identify random variables which could have their entropy reduced if asked about. In such work, clarification requests have taken the form of yes/no questions about the properties of an object [5, 9, 14] or generic wh-questions (e.g., "What do the words X refer to?") [22, 14].

Recent experimental evidence [11] suggests, however, that in HRI contexts, people prefer robots to list multiple options rather than asking for confirmation about a single referent with a yes/no- or generic wh-question (c.f. [3]). This is particularly striking as the evidence suggests that people maintain this preference even when a yes/no- or generic wh-question would be more efficient (c.f. [9]).

In contrast, Kruijff et al. present an approach in which robots can generate multiple-option clarification requests such as "Do you mean the blue or the red mug, Anne?" through a *continual planning* approach [10]. This approach, however, does not appear to be able to account for social context, uncertainty, or ignorance, and is only used for generation. The ability to handle social context is crucial for enabling natural HRI, and typical HRI scenarios are plagued by uncertainty and ignorance. An eldercare robot, for example, is not likely to be familiar with every object in the home of the elder it is assisting, nor with every person who might be referred to. Furthermore, the robot is unlikely to have *uncertainty-free* knowledge of all of the properties and relations involving those entities it *does* know of. We desire an approach that can account for these missing factors, and which can be used for both generation *and understanding*.

B. Design Hypotheses

In developing a new HRI-oriented approach to clarification request generation, our primary goal is to account for these missing factors. But we believe it is equally important to take *human preferences* into account as part of the design process. We believe that the previous work discussed thus far has not adequately considered what type of utterances humans *prefer* to use and be used. We hypothesize that there are three categories of human preferences that should affect the design decisions made when developing HRI-oriented clarification request generation algorithms.

Presentation of Options: Marge and Rudnicky (2015) suggests that people prefer that robots list options rather than ask yes/no- or generic wh-questions. But clearly there are limits to this preference. If a robot is asked "Could you get me some ice cream?" It is unlikely that humans will prefer a robot that lists twenty-seven available flavors instead of just asking "Which flavor would you like?" It is not yet clear, however, how many options can be listed until the use of a list is no longer preferable. We hypothesize **(H1)** that humans prefer options to be listed *only for a very small number of options*.

Demonstration of Intention Understanding: Similarly, many previous approaches use clarification requests that do not demonstrate understanding of the *meaning* of the sentence. If a robot is asked "Could you get me some ice cream," a robot that replies "What do the words 'ice cream' refer to" or "Do you mean 'the chocolate ice cream' or 'the vanilla ice cream"' does not allow its interlocutor to discern whether their *intention* was understood. In contrast, a robot that replies "*Would you like me to get you* the chocolate ice cream or the vanilla ice cream?" communicates understanding that the human wants ice cream *brought to them*. We hypothesize **(H2)** that humans prefer clarification requests that demonstrate understanding of their intentions.

Pragmatic Appropriateness: Finally, a robot that *does* generate clarification requests reflecting its understanding of human intentions will almost certainly need to use *indirect speech acts* [19] (e.g., *Would you like* me to get you the chocolate ice cream or the vanilla ice cream?), as the direct alternatives (e.g., "I have an intention to know whether you want me to have a goal to bring you the chocolate ice cream or the vanilla ice cream") are hard to express succinctly, and are viewed as less polite. We hypothesize **(H3)** that humans prefer indirectly rather than directly phrased clarification requests.

III. Experiment One: Preference Assessment

In this section, we present the results of a human subjects experiment designed to test our three hypotheses.

A. Methodology

Participants were recruited (20 Male, 10 Female) using Amazon Mechanical Turk. Participants ranged in age from 24 to 48 (M=32.67,SD=6.30). Each participant was asked seven simple questions, presented in a randomized order. Participants were told to imagine commanding a robot to "pick up the mug" in a scenario with several different-colored mugs on a table. For each question (which differed in the number of candidate mugs) two ways of asking for clarification were presented. Participants were asked to indicate which option they would prefer the robot to use.

The first five questions evaluated **H1**. In each case, participants chose between an option that listed out all options (ranging from "Would you like the red mug or the orange mug?" to "Would you like the red mug or the orange mug or the yellow mug or the green mug or the blue mug or the purple mug?") and a catch-all ("Which mug would you like?").

The sixth question evaluated **H2**. Participants chose between an option that indicated understanding of the speaker's goals ("Would you like the red mug or the green mug?") and one that did not ("Do you mean the red mug or the green mug?").

The last question evaluated **H3**. Participants chose between a pragmatically appropriate option ("Would you like the red mug or the blue mug?") and a pragmatically inappropriate option ("I have an intention to know if you want me to have a goal to bring you the red mug or the blue mug.")[1].

B. Results

As shown in Fig. 1, our results show that 70% of participants preferred options to be listed when there were only two options. But for more than two options, this number rapidly shrank. Only 20% of participants preferred options to be listed

[1] While this may seem a tortured construction, it is actually a straightforward verbalization of the type of logical expression commonly needed to be expressed in our robot architecture, and emphasizes how difficult it is to phrase clarification requests without using any sort of indirect language. In future work, however, it would be interesting to examine the effects of utterances that trade off pragmatic appropriateness for specificity in various ways.

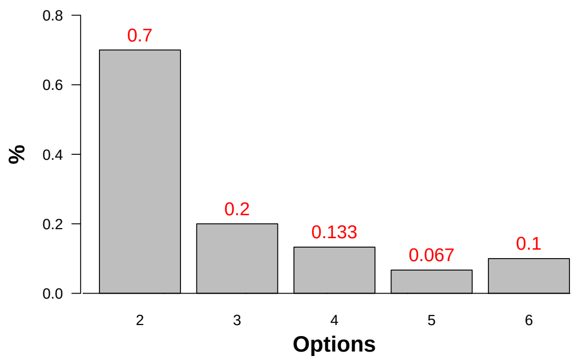

Fig. 1: **Experiment One Results**. Percentage of participants who preferred options to be listed, for each candidate number of options.

when there were three options, and preference for listing all options fell lower still when more options were listed. This confirms but clarifies the previous findings of Marge et al. [11], and suggests that robots likely do not need mechanisms for listing more than two options when there is referential ambiguity **(H1)**. Our results show that 80% of participants preferred the option that indicated understanding of their goals, supporting **H2**. Our results show that 93% of participants preferred the pragmatically appropriate option, supporting **H3**.

C. Discussion

The results of this experiment suggest three design recommendations. **(D1)** When phrasing clarification requests, if there are only two options, robots should present both options. Otherwise, robots should use a yes/no- or generic wh-question[2]. **(D2)** When phrasing clarification requests, robots should use phrasings that indicate that they understand the goals of their interlocutors. **(D3)** When phrasing clarification requests, robots should use pragmatically appropriate phrasings.

In the next section, we demonstrate how the integration of architectural components for reference resolution and pragmatic reasoning facilitates an approach to clarification request generation that not only fulfills all three of these design decisions, but also satisfies capabilities missing from previous approaches (e.g., context sensitivity, handling of uncertainty and ignorance, and use for both understanding and generation).

IV. Approach

In this section, we describe how each stage of our HRI-oriented clarification request generation framework [26] is handled by components of the DIARC architecture [18].

A. Uncertainty Identification

The first step in our clarification request generation framework is identifying whether or not there is uncertainty that needs to be clarified. To achieve this, we first determine the set of referential candidates and their respective levels of uncertainty. We then provide those candidates to a pragmatic inference component which produces a set of uncertain candidate interpretations. In this section, we will detail this process and the integration challenges it presents.

Notation (c.f.[21])

M	A robot's *world model* of entities $\{m_0 \dots m_n\}$.
Λ	A set of logical formulae $\lambda_0 \dots \lambda_n$, denoting (literal, direct) semantic *connotation* of an incoming utterance.
V	A set of free variables found in Λ.
Γ	A set of bindings from variables in V to entities in M, denoting the semantic *denotation* of an incoming utterance.
Φ	A *satisfaction* variable which is *True* iff all formulae in Λ *hold* when bound using Γ.

Our approach uses the DIST-POWER framework to facilitate access to information about entities a robot knows of [25]. The DIST-POWER framework uses a set of "consultants" to integrate a central, domain-independent open-world reference resolution component with a set of heterogeneous knowledge bases distributed throughout a robot architecture, potentially residing on multiple machines. In our instantiation of this framework, we make use of GH-POWER: our *Givenness Hierarchy*-theoretic reference resolution algorithm [28]. Based on the theoretical linguistic framework presented by Gundel et al. [8], GH-POWER treats DIST-POWER's distributed memory system as a Long Term Memory Store, and builds on top of it a set of hierarchical caches representing models of the robot's Discourse Context, Short-Term Memory, and Focus of Attention. This allows GH-POWER to resolve a wide array of referring expressions (REs). And, like the non-GH-theoretic version of POWER, GH-POWER handles both uncertain and open worlds. For the sake of simplicity, we will use POWER to refer to the distributed, GH-theoretic form of the POWER algorithm and its associated data structures.

POWER uses the logical form of an RE to (1) hypothesize new representations for previously unknown referents, and (2) produce a distribution $P(\Phi \mid \Gamma, \Lambda)$; that is, the probability of successful satisfaction conditioned on binding hypotheses from variables to *known* referents:

$$\{\Gamma_0 = \{\gamma_{0_0} \dots \gamma_{0_n}\} \dots, \Gamma_m = \{\gamma_{m_0} \dots \gamma_{m_n}\}\}$$

and semantic parse hypotheses:

$$\{\Lambda_0 = \{\lambda_{0_0} \dots \lambda_{0_n}\} \dots, \Lambda_m = \{\lambda_{m_0} \dots \lambda_{m_n}\}\}.$$

For example, suppose Bob asked Cindy "Can you grab the medical kit?" Cindy may parse this into something like

$$QuestionYN(b, s, can(s, grab(s, X)))$$

with additional semantic content $\Lambda_i = \{medkit(X)\}$ (Hereafter, we will use the abbreviations b="bob" and s="self"). If Cindy is 70% sure that the m_5 is a medical kit, reference resolution will produce:

$$P(\Phi = True \mid \Gamma = \{X \rightarrow m_5\}, \Lambda = \{medkit(X)\}) = 0.7$$

All sufficiently probable referential hypotheses are then used to create a set of *bound utterances with supplemental semantics* (BUSSes) $\Psi = \{\psi_0 \dots \psi_n\}$. Each $\psi_i \in$

[2]Future research will be needed to determine how the *content* of the options to be offered may impact how this decision is made. The results of such research may suggest refinements of this recommendation.

Ψ is associated with a unique sufficiently probable binding γ_i from variables found in the parsed utterance form and its supplemental semantics to entities found in Long Term Memory. For example, the BUSS associated with form $QuestionYN(b, s, can(s, grab(s, X)))$, semantics $\{medkit(X)\}$, and binding $\{X \rightarrow m_5\}$ would be:

$$\{QuestionYN(b, s, can(s, grab(s, m_5))) \wedge medkit(m_5)\}.$$

One could then create a *distribution* over BUSSes $P(\psi_i) = P(\Gamma_i, \Lambda_i \mid \Phi_i)$ using Bayes' Rule, *if* the next natural language component used a Bayesian framework. In fact, the next component in our architecture (i.e., the pragmatic reasoning component) uses a more general DS-theoretic framework [27], and thus another approach must be taken.

Dempster-Shafer (DS) Theory is a generalization of the Bayesian uncertainty framework that allows for elegant reasoning about uncertainty and ignorance even when distributional information is unavailable [20][3]. DS Theory is an attractive option for HRI domains in which agents may encounter new entities and concepts only a small number of times, with no information regarding the distribution underlying their occurrence. DS Theory is also useful for tasks such as pragmatic reasoning, because it would be impractical to store priors over all combinations of intentions and contexts, as would be required in a Bayesian framework; and because it allows the use of DS-based logical rules such as Modus Ponens, which cannot be used in a strictly Bayesian framework.

But not all components of a robot's architecture will likely be DS-theoretic. For some components, distributional information may well be available, allowing for use of a Bayesian approach. Each architectural component should be able to use the knowledge representation and uncertainty management approach most conducive to its own operation. To facilitate this, researchers have developed mechanisms that allow seamless integration between components with different uncertainty management schemes [26].

Using the mechanisms discussed in such work, we produce a DS-theoretic *Frame of Discernment (FoD)* Θ of hypotheses described by the logical conjunctions (i.e., BUSSes) $\{\psi_0 \dots \psi_n\}$. Remember that each BUSS contains both a parsed utterance form and a set of supplemental semantics, bound using a single candidate variable binding. The next component in the DIARC NL Pipeline (i.e., PRAG) only uses the utterance form, however, and there may be multiple hypotheses in the resulting Frame of Discernment Θ that have the same utterance form but different supplemental semantics. Note, however, that each ψ only uses those bindings in Γ_i associated with the utterance's root node (typically the formula representing the verb). There may be variables in V that had multiple possible bindings, but which do not appear in the utterance's root node, and thus there may be identical hypotheses within our frame of discernment.

For example, if Bob had asked "Can you grab the medkit that is near the book?", and one candidate medkit (m_1) is actually near two books (m_2 and m_3), we could have two hypotheses which can be described by BUSSes that have the same utterance form (e.g. $QuestionYN(b, s, grab(s, m_1))$) but different supplemental semantics (e.g., $\{medkit(m_1) \wedge book(m_2) \wedge near(m_1, m_2)\}$ vs $\{medkit(m_1) \wedge book(m_3) \wedge near(m_1, m_3)\}$. We thus cluster these hypotheses into sets $C_0, \dots, C_n$ such that all hypotheses associated with each set are described by BUSSes that have the same utterance form. For example, if we have three singleton hypotheses $\{\theta_1, \theta_2\, \theta_3\}$, and ψ_1 and ψ_2 have the same utterance form, $C = \{\{\theta_1, \theta_2\}, \{\theta_3\}\}$.

We can now split our FoD Θ into a set of $\mid C \mid$ "binary" FoDs, one for each cluster C_i. Each binary FoD itself has two hypotheses: (1) that the utterance form describing all hypotheses in cluster C_i represents what was communicated, and (2) that it does not. This splitting has no theoretical ramifications, but facilitates easier integration with PRAG. Because each cluster is mutually exclusive from all other clusters, each binary FoD can be represented entirely by the *bound utterance structure*:

$$\langle utterance(\psi_i), Bl(\{C_{i_0} \dots C_{i_n}\}), Pl(\{C_{i_0} \dots C_{i_n}\})\rangle.$$

Suppose $\Theta = \{\theta_1, \theta_2, \theta_3\}$ and $\Psi = \{\psi_1, \psi_2, \psi_3\}$, where

$$\begin{aligned}
\psi_1 =& (QuestionYN(b, s, can(s, grab(s, m_1))) \\
& \wedge medkit(m_1) \wedge book(m_2) \wedge near(m_1, m_2)), \\
\psi_2 =& (QuestionYN(b, s, can(s, grab(s, m_1))) \\
& \wedge medkit(m_1) \wedge book(m_3) \wedge near(m_1, m_3)), \\
\psi_3 =& (QuestionYN(b, s, can(s, grab(s, m_4))) \\
& \wedge medkit(m_4) \wedge book(m_2) \wedge near(m_4, m_2)),
\end{aligned}$$

and assume the following DS-theoretic *Basic Belief Assignment(BBA)* assigning probability masses to each hypothesis in Θ, where Bl and Pl (belief and plausibility) are upper and lower bounds on the expected probability of each hypothesis:

Hypothesis	Mass	Bl	Pl
$\emptyset$	0.0	0.0	0.0
$\{\theta_1\}$	0.2	0.2	0.2
$\{\theta_2\}$	0.3	0.3	0.3
$\{\theta_3\}$	0.5	0.5	0.5
$\{\theta_1, \theta_2\}$	0.0	0.5	0.5
$\{\theta_2, \theta_3\}$	0.0	0.8	0.8
$\{\theta_3, \theta_1\}$	0.0	0.7	0.7
$\{\theta_1, \theta_2, \theta_3\}$	0.0	1.0	1.0

Because ψ_1 and ψ_2 have the same utterance form, $C = \{\{\theta_1, \theta_2\}, \{\theta_3\}\}$. From this, the following set of bound utterance structures will be created:

$$\begin{aligned}
&\{\langle QuestionYN(b, s, can(s, grab(s, o_1))), \\
&\qquad Bl(\{\theta_1, \theta_2\}), Pl(\{\theta_1, \theta_2\})\rangle, \\
&\langle QuestionYN(b, s, can(s, grab(s, o_4))), \\
&\qquad Bl(\{\theta_3\}), Pl(\{\theta_3\})\rangle\} = \\
&\{\langle QuestionYN(b, s, can(s, grab(s, o_1))), 0.5, 0.5\rangle \\
&\langle QuestionYN(b, s, can(s, grab(s, o_4))), 0.5, 0.5\rangle\}
\end{aligned}$$

The set of bound utterance structures is sent to PRAG, which uses context to determine the intentions underlying utterances [27], producing a set of intentional structures

[3]For reasons of space, this paper does not include a treatment of the basic notions of DS Theory. We direct the unfamiliar reader to [27], upon which this work builds and which covers DS Theory's basic notions, or to [6], which provides a more thorough treatment for the uninitiated.

$\langle I, Bl(I), Pl(I)\rangle$. If the difference between $Bl(I)$ and $Pl(I)$ is sufficiently large, or if $\frac{Pl(I)+Bl(I)}{2}$ is sufficiently close to 0.5, (assessed using Nunez' uncertainty measure [13]), intention I is deemed in need of clarification. PRAG then formulates an intention-to-know (*itk*) which of these intentions is correct, denoted $itk(s, or(i_0, i_1, \ldots, i_n))$.

Before integration with POWER, PRAG only handled *pragmatic* uncertainty. Because PRAG now receives a set of candidate utterance forms with potentially different argument bindings, it now also automatically handles *referential* uncertainty.

Before we move on, we would like to point out that that because DIARC's reference resolution component handles *open worlds*, instances in which interlocutors refer to previously unknown entities do not automatically generate clarification requests. For example, if the robot is told "Go to the room at the end of the hall" and does not know of a room at the end of the hall, it will not ask for clarification, but will rather hypothesize a new location, and carry on. We do not regard such situations as referentially ambiguous (although it may be valuable to ask for more information about this location). Here, the robot knows what entity is being referred to: a previously unknown room at the end of the hall.

B. Decision to Communicate

Currently, any *intention-to-know (itk)* formulated during the previous stages is automatically asserted into the robot's knowledge base, triggering a decision to communicate this intention once it is acceptable for the robot to accept the conversational turn. When this decision is made, the *itk* is passed to the pragmatic *generation* component for processing.

C. Utterance Choice

The robot must now determine a contextually appropriate way to formulate its intention at the *utterance level*. This is accomplished once again by PRAG, which uses the same set of rules for generation as it uses for inference [27]. In Experiment One, we observed that if there were more than two options, listing those options was dispreferred over a more general question. Thus, if we are to send a clarification request to PRAG that has semantics of the form $itk(self, or(option_1, ...option_n))$, we first check whether or not n is greater than the acceptable number of candidates to list, i.e., two. If $n = 2$, this intention is sent directly to PRAG. Otherwise, all options are unified into a single predicate whose only bound arguments are those that are identical for all options. For example, if $\{option_1, option_2, option_3\}$ = $\{need(jim, obj_1), need(jim, obj_2), need(jim, obj_3)\}$, these will be unified into $need(jim, ?)$, and the intention $itk(self, need(jim, ?))$ will be sent to PRAG instead.

Using DS-theoretic logical operators, PRAG determines a set of candidate utterance forms, each of which is forward-simulated through pragmatic inference to ensure that the agent does not accidentally communicate anything it does not actually believe to be true as a side effect of communicating its primary illocutionary point. The best candidate utterance is then sent to NLG for surface realization.

This processing step is not typically included in traditional NLG frameworks, which do not typically need to account for social context or dialogue context. They instead typically include a *document structuring* (c.f. [16]) stage in which the agent determines the order in which to convey multiple utterances. Because clarification request generation in HRI *typically* only involves a single utterance, we do not currently handle this step, but it will be an important topic for future work. A robot may, for example, need to preface a clarification request by stating what parts of an utterance it *did* understand.

D. Surface Realization and Speech Synthesis

Once an appropriately phrased utterance form is chosen by the pragmatic generation component, that utterance is sent to the *Natural Language Generation* component for Surface Realization. First, that component chooses sets of properties to use to describe each of the utterance's referents. For example, consider the utterance form

$$QuestionWH(s, b, or(need(b, grab(s, m_1)), need(b, grab(s, m_2)))).$$

Here, there are two referents that must be described: m_1 and m_2. The referent m_1 may be described using the properties $\{mug(X) \wedge white(X)\}$, and m_2 may be described using the properties $\{mug(X) \wedge black(X) \wedge large(X)\}$.

The utterance form and sets of properties are then translated into raw text using the open source SimpleNLG package, producing, for example, "Do you need the white mug or do you need the large black mug?" when there are two referential candidates, and producing, for example, "Which one do you need?" in the case of a larger number of referential candidates. The open source MaryTTS package is then used to synthesize this text into an audio form that is produced by the robot.

V. DEMONSTRATION

To demonstrate the operation of the presented approach, we present a proof-of-concept interaction that occurs in a simulated environment. This demonstration highlights the full implementation of all stages of the clarification request generation framework through components of the DIARC architecture. Specifically, this demonstration uses the components of the DIARC architecture shown in Fig. 2. In addition to components responsible for the simulation of a Pioneer robot within an office environment, our configuration used the following components: ASR (which performs simulated speech recognition), NLP (which uses the C&C parser within a GH-theoretic framework), POWER (which performs reference resolution), AGENTS, SPEX and OBJECTS (POWER Consultants (c.f. [24]) providing information about people, places, and things), DIALOGUE (which, performs dialogue management, and includes PRAG as a submodule), BELIEF (which allows DIALOGUE to assess its current context), and ACTION (which performs goal and action management). Of these components, the POWER, NLP, NLG, and DIALOGUE components are central to the integrated approach presented in this paper.

The interaction begins with the speaker saying to the robot "I need the medkit" in an environment in which the robot knows of two medkits, one red and one white. ASR sends this sentence to NLP, which parses the utterance into a

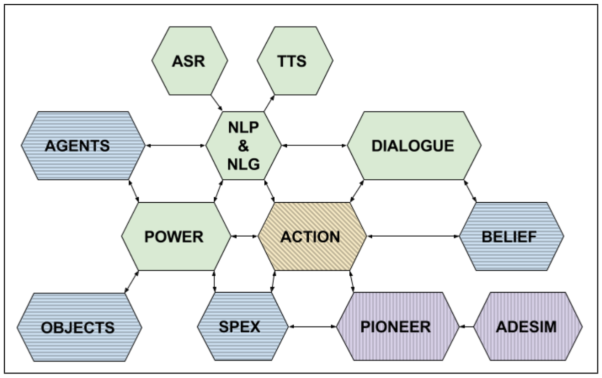

Fig. 2: **Architecture Diagram**. Knowledge base components are depicted in blue (horizontal stripes), linguistic components in green (no stripes), simulation components in purple (vertical stripes)), and the action manager in yellow (diagonal stripes).

dependency tree, from which it extracts root semantic content $need(X1, X2)$, with utterance type $Statement$, additional semantic content $\{speaker(X1) \wedge medkit(X2)\}$, and presumed cognitive statuses $\{X1 \rightarrow definite, X2 \rightarrow definite\}$. Using this information, POWER searches for the referents to bind to $X1$ and $X2$; for $X1$, POWER finds a single probable candidate: agt_1, with probability 1.0; for $X2$, two candidates are found: obj_1, with probability of satisfaction 0.82, and obj_2, with probability of satisfaction 0.92. These bindings are then used to create the following bound utterances[4]:

$$\{Stmt(b, s, need(b, obj_1)), \quad Stmt(b, s, need(b, obj_2))\}$$

with corresponding probabilities[5] 0.82 and 0.92, respectively. These are normalized and used to create DS-theoretic bound utterance structures, which are passed to DIALOGUE:

$$\{\langle Stmt(b, s, need(b, obj_1)), 0.471, 0.471\rangle, \\ \langle Stmt(b, s, need(b, obj_2)), 0.529, 0.529\rangle\}$$

PRAG possesses the rule:

$$\langle Stmt(X, Y, need(Z, W)) \Rightarrow goal(Y, bring(Y, W, Z)), 0.9, 0.99\rangle, \tag{1}$$

indicating that the robot is between 90 and 99% confident in the rule; because the antecedent of this rule matches the utterance form of each bound utterance structure, uncertain Modus Ponens is applied in both cases, producing the set of intentional structures:

$$\{\langle goal(s, bring(s, obj_1, b)), 0.424, 0.576\rangle, \\ \langle goal(s, bring(s, obj_2, b)), 0.476, 0.524\rangle\}$$

Note that at this point, belief no longer equals plausibility: while the robot may not have encoded any ignorance with respect to what utterance was heard, ignorance encoded with respect to the context and rules the robot uses for pragmatic inference are reflected in the uncertainty intervals of the rules' consequents, thus painting a better picture of how much the robot truly knows about its interlocutor's intentions.

Nunez' uncertainty rule determines that both of these intentions are highly uncertain. DIALOGUE thus determines its own intention to know which is correct, encoded as the structure:

$$\langle itk(s, or(goal(s, bring(s, obj_1, b)), \\ goal(s, bring(s, obj_2, b)))), 1.0, 1.0\rangle$$

To decide how to communicate this intention, the bound utterance structure is passed through PRAG in reverse [27], using a rule of the form:

$$\langle QuestionWH(X, Y, or(Z, W)) \Rightarrow \\ itk(X, or(Z, W)), 0.95, 0.95\rangle, \tag{2}$$

Our approach allows recursive generation, and thus Eq. 2 is chained with Eq. 1 to produce:

$$QuestionWH(s, b, or(need(b, obj_1), need(b, obj_2))).$$

This utterance is then sent to our NLG component for generation of REs for "bob", "obj_1" and "obj_2", and subsequent realization of the entire expression. This produces the text "Do you need the white medkit or do you need the red medkit?" which is then synthesized and output by the robot.

VI. Experiment Two: Human-Subjects Evaluation

To evaluate our approach, we conducted a human-subject experiment similar to Experiment One, comprised of (1) a data collection stage, and (2) an evaluation stage.

A. Data Collection

We first created a tabletop scene containing twelve objects: four different colored waterbottles, four different colored markers, and four different colored mugs (Fig. 3). For each object type, we took photographs of the scene in which zero, one, or two of that object type were taken away. This produced nine tabletop scenes, three of which contained identical object arrangements (i.e., those scenes in which no objects were removed). In our data collection experiment, each participant was shown one of these nine images at random, with a caption describing the participant's task, followed by a text box. For example, for the image in which three of the four waterbottles was shown, the following caption was used:

"You have been told 'I need the bottle!' and would like to fulfill the speaker's request. However, as you can see, there are three bottles on the table: a silver bottle, a green bottle, and a blue bottle. Please type a sentence you would use to ask the speaker for clarification, so that you will know what bottle to pick up."

Fig. 3: Tabletop Environment used in Experiment Two.

[4]Here, agt_1 is changed to the agent's name for dialogue processing.

[5]All beliefs and plausibilities in this section are rounded.

Vocabulary	Types	Type/Token Ratio	Diversity [2, 17]
596	64	.107	1.85

TABLE I: Vocabulary statistics for utterances collected in Experiment Two, Part One.

Similar captions were used for the other images. Once the participant entered text into the text box, they were free to click to the next page, and end the experiment.

Participants were recruited (53 Male, 39 Female) using Amazon Mechanical Turk. Participants ranged in age from 20 to 77 (M=33.15,SD=8.94), and were paid $0.30 to participate. As a total of 92 participants were recruited, an average of 30.7 utterances were collected for each grouping of scenes that had the same number of objects removed. Vocabulary diversity statistics for these utterances are reported in Tab. I.

All utterances collected in this stage were standardized with respect to noun phrasing. For example, "Do you want me to pick up the silver bottle or the blue bottle?" was reduced to "Do _ want _ to pick up _ or _?" All utterances within each cluster were grouped by identical phrasing, and the three most common phrasings for each cluster were selected (four in the case of a tie). The REG algorithm described above was then used to generate noun phrases to fill into the previously created gaps, thus creating three to four utterances for each image.

Next, an additional utterance was generated for each image using the approach presented in this paper: for each image, knowledge of the objects in the image was provided to the robot architecture, and the utterance "I need the [name of object type]" was said to a robot running the architecture. Because the architecture also used the REG algorithm described above, the utterances generated by our robot architecture had the same noun-level phrasings as all other utterances, but a different utterance-level phrasing. Thus, this stage produced a set of thirty-nine utterances with unique utterance level phrasings but identical noun level phrasings. The thirteen utterance forms (before REs were filled in) are shown in Tab. II, Column 3.

B. Evaluation

In this stage, each participant was shown one of the nine tabletop scenes created in the first stage, along with a caption such as: "Your friend Alex says to you, 'I need the bottle!' Which of the following sentences would be best to say to Alex, so that you will know which bottle to give her?"

Each participant was then presented with the four to five utterances associated with the presented image, in the form of buttons. Clicking on one of the utterances moved the participant to the next page, and ended the experiment. Participants were recruited (94 Male, 88 Female) using Amazon Mechanical Turk. Participants ranged in age from 18 to 74 (M=34.55, SD=11.16), and were paid $0.30 to participate. As a total of 182 participants were recruited, an average of 20.22 data points were collected for each scene.

Robot-generated requests were chosen only slightly less frequently than were human-generated requests: overall, robot-generated requests were chosen 18.13% of the time, whereas each form of human-generated request was chosen, on average, 24.67% of the time. Overall, this is a positive result as it suggests that the algorithm overall did not generate requests that were much worse than the requests that humans used most frequently. A request-by-request breakdown of participants' choices is shown in Tab. II, Column 4.

Generator	#	Utterance Generated in Part One	Result
Robot	2	Do you need __ or do you need __?	9.4%
Human	2	Do you need __ or __?	**45.3%**
Human	2	What color __ do you need?	22.6%
Human	2	What color __ do you want?	22.6%
Robot	3	Which one do you need?	23.7%
Human	3	Which color __ do you need?	**33.9%**
Human	3	Which color __?	23.7%
Human	3	Which color __ would you like?	18.6%
Robot	4	Which one do you need?	20.0%
Human	4	What color __ do you need?	**24.3%**
Human	4	Which color __ would you like?	22.9%
Human	4	Which color __?	21.4%
Human	4	What color is the __?	11.4%

TABLE II: Utterance forms generated in Experiment Two, Part One, and chosen between in Experiment Two, Part Two. *Col. 1* indicates whether each utterance form was generated by the presented approach or by a human in Part One. *Col. 2* indicates how many suitable referents existed in the scene for which each utterance was generated. *Col. 3* indicates the generated utterance form, generalized across noun phrases. In Part Two, blanks were filled with generated REs. For example, in scenes with initial utterance "I need the bottle", gaps in the first two rows were filled with "the green bottle" and "the silver bottle", and remaining gaps were filled with "bottle". *Col. 4* indicates the percentage of participants in Part Two who chose that utterance form as the best to use to ask for clarification.

But in fact, robot-generated requests stand to perform significantly *better* than the majority of human-generated requests when there are exactly two options to disambiguate. As shown in the first section of Tab. II, in this case our robot-generated requests were chosen significantly less frequently than were human-generated requests, but were nearly identical to the top performing human-generated requests. The robot-generated requests were simply more verbose, as they used a conjunction at the clause level rather than the noun-phrase level. This suggests that if our approach had been modified to use conjunctions at the phrase level, it may have outperformed the second- and third-best human-generated requests *combined*. We have since made this modification, as we will later discuss.

C. Discussion

In Experiment One, we observed that participants dispreferred clarification requests that were insensitive to pragmatic factors, did not indicate understanding of an interlocutor's goals or intentions, listed more than two options, or did not list both options when there were only two likely candidates. These observations were confirmed in Experiment two, part two. The most commonly chosen clarification requests were nearly identical to the clarification requests generated by our robot architecture. But in neither the two-, three-, or four-option utterance groupings were our chosen clarification requests *exactly* identical to the most commonly chosen clarification requests, and in fact differed from those requests in small but important ways.

As previously mentioned, when there was referential ambiguity between only two candidate referents, participants in Experiment Two Part Two preferred clarification requests that listed all options. However, the specific phrasing used by our robot architecture was simply too verbose, as it failed to identify structural similarities and distribute appropriately. Since running this evaluation, we have added functionality to the NLG component that performs such distribution when structural similarity is detected, and our robot architecture thus now generates the exact utterance forms that were most preferred by humans (e.g., "Do you need __ or __" rather than "Do you need __ or do you need __?"). Future work will be needed to determine the distribution of selections that would be seen if the overly verbose (originally robot-generated) RE were not presented. We would expect the most common human-generated (and now, robot-generated) RE to be chosen between 45.3 and 54.5% of the time, putting the robot's performance on par with human performance.

A greater difference is observed when more than two options present themselves. It is striking to observe that all commonly-used human-generated utterances in these cases do not explicitly ask for disambiguation between bottles, but rather ask for information regarding a specific property that could be used to disambiguate between bottles. This suggests that in these cases, the optimal approach to clarification request generation likely lies somewhere between the approach presented in this paper and the information-theoretic approaches seen in previous work [5, 9, 14]. We predict that the ideal approach to clarification request generation may involve generation in a way quite similar to the approach used in this paper, followed by a stage in which information-theoretic mechanisms are used to add differentiating modifiers.

It is also important to note, however, that in all three cases a significant percentage of participants did choose the less popular choices. When four options were presented, for example, "Which color__ would you like" was chosen by less than two percent fewer participants than was the most popular "What color __ do you need?". This suggests that it may be valuable in future work to develop models of human interlocutors that model this type of individual difference.

While at first glance the difference between the alternate strategies may seem arbitrary, we suspect that they in fact represent different strategies that are either explicitly used, or which arise from differential weightings of pragmatic principles. Utterances such as "Which color __ *do you need*" may be used due to subconscious *lexical entrainment* or conscious *refashioning* in which speakers use the same phrasing as that used by their interlocutors [4, 1, 29]. Utterances such as "Which color __ *would you like*" and "Which color __ *do you want*" may be used if the pragmatic value of a refashioned sentence is weighted lower than that of a more *conventionally indirect* utterance form [19]. And utterances such as "Which color_,' may be used due to the interaction of either aforementioned pragmatic strategy with Grice's Third Maxim of Manner: "Be brief (avoid unnecessary prolixity)" [7].

VII. Conclusion

We have presented an integrated approach to clarification request generation for HRI contexts, and shown how this approach can identify and handle both pragmatic and referential ambiguity. We have also shown how our approach can be used in architectures where information about referents is distributed across multiple heterogeneous knowledge bases, as is often the case in cognitive robot architectures. The primary finding of this paper is that a language-enabled robot's pragmatic reasoning component can track and address *referential* ambiguity when integrated with a probabilistic reference resolution component: a useful finding for designers of language-enabled robot architectures intended for use in HRI domains. We furthermore demonstrated an implementation of this approach on a simulated robot.

In addition, we have provided the results of two human-subject studies. Our first study replicated and refined the recommendations of previous studies of human-robot dialogue. Our second experiment showed that the theoretical commitments of our robot architecture align with human preferences, and that the clarification requests generated by our full NLG pipeline may be comparable to human-generated clarification requests.

Our findings suggest several directions for future work. First, research is needed on using information-theoretic mechanisms to adapt clarification requests generated by pragmatic reasoning components. Second, research is needed to develop speaker-specific models that can predict precisely what type of clarification request they would most likely prefer, based on their inferred weighting of pragmatic principles. Third, future work should also further examine methods by which components using different frameworks for representing uncertainty can be optimally integrated. Finally, a tighter integration between pragmatic reasoning and reference resolution can be achieved. In previous work, we have shown how our pragmatic reasoning component can use contextual knowledge to abduce the most appropriate way to phrase an utterance; but this contextual knowledge is assumed to be stored in a robot's centralized belief and dialogue components. In future work, this should be extended to allow this knowledge to be appropriately distributed across the robot's heterogeneous knowledge bases, as is its other knowledge.

VIII. Acknowledgments

This work was in part funded by grant N00014-14-1-0149 from the US Office of Naval Research.

References

[1] Susan E Brennan, Alexia Galati, and Anna K Kuhlen. Two minds, one dialog: Coordinating speaking and understanding. *Psychology of Learning and Motivation*, 53: 301–344, 2010.

[2] John B Carroll. *Language and thought*. Foundations of Modern Psychology. Prentice-Hall, Englewood Cliffs, N.J., 1964.

[3] Herbert H Clark. *Using language*. Cambridge University Press, 1996.

[4] Herbert H Clark and Edward F Schaefer. Contributing to discourse. *Cognitive Science*, 13(2):259–294, 1989.

[5] Robin Deits, Stefanie Tellex, Thomas Kollar, and Nicholas Roy. Clarifying commands with information-theoretic human-robot dialog. *Journal of Human-Robot Interaction (JHRI)*, 2(2):58–79, 2013.

[6] Jean Gordon and Edward H Shortliffe. The Dempster-Shafer theory of evidence. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, 3:832–838, 1984.

[7] Herbert P Grice. Logic and conversation. *Syntax and Semantics*, 3:41–58, 1970.

[8] Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307, 1993.

[9] Sachithra Hemachandra, Matthew R Walter, and Seth Teller. Information theoretic question asking to improve spatial semantic representations. In *Proceedings of the AAAI Fall Symposium Series*, 2014.

[10] Geert-Jan M Kruijff, Michael Brenner, and Nick Hawes. Continual planning for cross-modal situated clarification in human-robot interaction. In *Proceedings of the Seventeenth IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 592–597, 2008.

[11] Matthew Marge and Alexander I Rudnicky. Miscommunication recovery in physically situated dialogue. In *Proceedings of the Sixteenth Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 22–49, 2015.

[12] Maja J Matarić. Situated robotics. *Encyclopedia of Cognitive Science*, 2002.

[13] Rafael C Núñez, Ranga Dabarera, Matthias Scheutz, Gordon Briggs, Otavio Bueno, Kamal Premaratne, and Manohar N. Murthi. DS-based uncertain implication rules for inference and fusion applications. In *Proceedings of the Sixteenth International Conference on Information Fusion (FUSION)*, pages 1934–1941, 2013.

[14] Matthew Purver. Clarie: The clarification engine. In *Proceedings of the eigth SEMDIAL Workshop on the Semantics and Pragmatics of Dialogue (CATALOG)*, pages 77–84, 2004.

[15] Matthew Purver, Jonathan Ginzburg, and Patrick Healey. On the means for clarification in dialogue. In *Current and New Directions in Discourse and Dialogue*, pages 235–255. Springer, 2003.

[16] Ehud Reiter, Robert Dale, and Zhiwei Feng. *Building natural language generation systems*. MIT Press, 2000.

[17] Brian Richards. Type/token ratios: What do they really tell us? *Journal of Child Language*, 14(02):201–209, 1987.

[18] Matthias Scheutz, Paul Schermerhorn, James Kramer, and David Anderson. First steps toward natural human-like HRI. *Autonomous Robots*, 22(4):411–423, 2007.

[19] John R Searle. Indirect speech acts. *Syntax and Semantics*, 3:59–82, 1975.

[20] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[21] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. Approaching the symbol grounding problem with probabilistic graphical models. *AI Magazine*, 32(4):64–76, 2011.

[22] Stefanie Tellex, Pratiksha Thaker, Robin Deits, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. Toward information theoretic human-robot dialog. *Robotics*, 32: 409–417, 2013.

[23] David R Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, Rochester, NY, 1994.

[24] Tom Williams and Matthias Scheutz. POWER: A domain-independent algorithm for probabilistic, open-world entity resolution. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1230–1235, 2015.

[25] Tom Williams and Matthias Scheutz. A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, pages 3598–3964, 2016.

[26] Tom Williams and Matthias Scheutz. Resolution of referential ambiguity using Dempster-Shafer theoretic pragmatics. In *Proceedings of the AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction (AI-HRI)*, 2016.

[27] Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. Going beyond command-based instructions: Extending robotic natural language interaction capabilities. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, pages 1387–1393, 2015.

[28] Tom Williams, Saurav Acharya, Stephanie Schreitter, and Matthias Scheutz. Situated open world reference resolution for human-robot dialogue. In *Proceedings of the Eleventh ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 311–318, 2016.

[29] Si On Yoon and Sarah Brown-Schmidt. Lexical differentiation in language production and comprehension. *Journal of Memory and Language*, 69(3):397–416, 2013.