Continual Processing of Situated Dialogue in Human-Robot Collaborative Activities

Geert-Jan M. Kruijff, Miroslav Janíček and **Pierre Lison**
Language Technology Lab
German Research Center for Artificial Intelligence, DFKI GmbH
{gj,miroslav.janicek,pierre.lison}@dfki.de

Abstract— **This paper presents an implemented approach of processing situated dialogue between a human and a robot. The focus is on task-oriented dialogue, set in the larger context of human-robot collaborative activity. The approach models understanding and production of dialogue to include intension (what is being talked about), intention (the goal of why something is being said), and attention (what is being focused on). These dimensions are directly construed in terms of assumptions and assertions on situated multi-agent belief models. The approach is continual in that it allows for interpretations to be dynamically retracted, revised, or deferred. This makes it possible to deal with the inherent asymmetry in how robots and humans tend to understand dialogue, and the world in which it is set. The approach has been fully implemented, and integrated into a cognitive robot. The paper discusses the implementation, and illustrates it in a collaborative learning setting.**

I. Introduction

Particularly in task-oriented dialogues between a human and a robot, there is more to dialogue than just understanding words. The robot needs to understand what is being talked about, but it also needs to understand why it was told something. In other words, what the human intends the robot to do with the information, in the larger context of their joint activity.

In this paper we see task-oriented dialogue as part of a larger collaborative activity, in which a human and the robot are involved. They are planning together, executing their plans. Dialogue plays a facilitatory role in this. It helps all participants build up a common ground, and maintain it as plans are executed, and the world around them changes.

We present here an approach that models these aspects of situated task-oriented dialogue. We provide an algorithm in which dialogue is understood, and generated, by looking at *why* something is being said (intention), *what* that something is about (intension), and *how* that helps to direct the focus (attention). Core to the algorithm is abductive reasoning. This type of reasoning tries to find the best explanation for observations. In our case, it tries to find the best explanation for why something was said (understanding), or how an intention best could be achieved communicatively (generation). Thereby, abduction directly works off the situated, multi-agent belief models the robot maintains as part of its understanding of the world, and of the agents acting therein.

Our approach views dialogue from a more intentional perspective, like the work by Grosz & Sidner [6], Lochbaum et al. [10], and most recently Stone et al [14], [15], [16]. Our approach extends that of Stone et al.

Stone et al. formulate an algorithm for collaborative activity, involving abductive reasoning. They assume that understanding and production are symmetric: "what I say is how you understand it". However, this is optimistic for human-human dialogue, and rather unrealistic for human-robot interaction. Robots hardly ever perfectly understand what is meant. We need to allow for the robot to act upon interpretations even when they are incomplete or uncertain. And, should it turn out that the robot has misunderstood what was said, roll dialogue back to a point where the robot can clarify and correct its understanding.

Our approach enables these features by introducing *assertions* into our logics. This idea is inspired by Brenner & Nebel's work on continual planning [3]. An assertion is a content formula that needs to be verified at a later point. In that, it is different from a propositional fact, which the robot knows to be either true or false. We can introduce an assertion into an abductive inference to help find an explanation, and then act upon it. It is just that this is then made contingent on the assertion to become true sooner or later. In this paper, we show how assertions can play a fundamental role in helping a robot and a human achieve common ground in collaborative activity.

Below, §II provides a brief overview on intentional approaches to dialogue. §III presents our approach and discusses situated multi-agent belief models, abductive reasoning, and the algorithm for continual processing of collaborative activity. §IV discusses the implementation, and §V illustrates it on working examples from an integrated robot system.

II. Background

Recent theories of dialogue focus on how participants can obtain common ground through alignment [11]. Agents align how they communicate content, what they pay attention to, and what they intend to do next. They base this on how they perceive each other's views on the world.

This works out reasonably well as long as we can assume a more or less common way of "looking" at things. Even when humans normally differ in what they know, can, and intend to do, there is typically a common categorical framework in which they can characterize the world, in order to arrive at a common ground. This is where a problem arises in communication between a human and a robot that continuously learns, because robots tend to see things substantially differently than humans. For this reason, mechanisms for modeling, and dealing with, such asymmetry in understanding are necessary for situated dialogue. We present here an approach providing such means.

The approach is based on an extension of Stone & Thomason's (S&T) abductive framework [14], [15], [16]. S&T model comprehension and production of dialogue as construction of abductive proofs. Abduction reasons towards an explanation consisting of a consistent context update and possible changes to attentional state. The explanation is based on factual assumptions, observations, and inferred intentions, all included at a context-sensitive cost. They thus place belief context, attention, and intention on par. This is similar to other intentional approaches to dialogue and discourse, such as Grosz & Sidner's [6]. S&T's approach arguably provides more flexibility [16] in that aspects such as reference resolution are dynamically determined through proof, rather than being constrained by hierarchical composition of a context model. For comprehension, an abductive proof provides the conditions under which an agent can update its belief model and attentional model with the content for a communicated utterance, and its task model using the inferred intentions underlying the utterance. For production, an abductive proof provides the conditions for executing a plan to achieve an intended context and attentional update in another agent.

We extend S&T in several ways. We expand context [14] to incorporate the types of situated multi-agent beliefs and tasks with which the robot reasons in understanding collaboration, and the world. We also make S&T's notion of "checkpoints" more explicit. A checkpoint is a means to establish whether assumptions are in fact warranted [16]. Checkpoints introduce a relation between constructing an explanation, and acting on it. This suggests a similarity to the construction of a plan and the monitoring of its execution. [3] introduces a notion of assertion for continual planning. An assertion poses the availability of future observations, to enable the construction of a continual plan including actions based on such an assertion. Upon execution, assertions are checked and are points for possible plan revision.

We propose to use a similar notion. In an abductive proof, we can include assumptions, observations, and actions at varying costs to infer an explanation. They all contribute facts or outcomes from which further inferences can be drawn. An assertion is a statement whose truth needs to assumed, but that cannot be proved or disproved on the current set of beliefs of the agent. Marking assertions turns these statements in an abductive proof into points that warrant explicit verification – i.e. they act as checkpoints. The notions of assertion and checkpoint provide our approach with a fundamental way of dealing with asymmetry in understanding, and resolving it to come to common ground.

III. Approach

A. Modeling multi-agent beliefs

We base our approach to situated grounding in direct reasoning about the agents' *beliefs*. A belief is an agent's informational state that reflects its understanding of the world and the way it has been talked about. Such an understanding can be acquired through direct observation (as a result of a sensoric input), or through communication with other agents, as is the case when engaging in a dialogue. Moreover, these beliefs can explicitly model *common beliefs*, which correspond to the beliefs that are a part of the common ground among a group of agents.

A belief is represented as a formula $\mathsf{B}e/\sigma : \phi$ that consists of three parts: a *content formula* ϕ from a *domain logic* $\mathcal{L}_{\mathsf{dom}}$, the assignment e of the content formula to agents, which we call an *epistemic status* and the *spatio-temporal frame* σ in which this assignment is valid.

We distinguish three classes of epistemic statuses, that give rise to three classes of beliefs:

- **private** belief of agent a, denoted $\{a\}$, comes from *within* the agent a, i.e. it is an interpretation of sensor output or a result of deliberation.
- a belief **attributed** by agent a to other agents $b_1, ..., b_n$, denoted $\{a[b_1, ..., b_n]\}$, is a result of a's deliberation about the mental states of $b_1, ..., b_n$ (e.g. an interpretation of an action that they performed).
- a belief **shared** by the group of agents $a_1, ..., a_m$, denoted $\{a_1, ..., a_m\}$, is common ground among them.

A spatio-temporal frame is a contiguous spatio-temporal interval. The belief is only valid in the spatio-temporal frame σ and frames that are subsumed by σ. In this way, spatio-temporal framing accounts for situatedness and the dynamics of the world. The underlying spatio-temporal structure may feature more complex spatial or temporal features.

Finally, the domain logic $\mathcal{L}_{\mathsf{dom}}$ is a propositional modal logic. We do not require $\mathcal{L}_{\mathsf{dom}}$ to have any specific form, except for it to be sound, complete and decidable.

Multiple beliefs form a *belief model*. A belief model is a tuple $\mathbf{B} = (A, S, K)$ where A is a set of agents, S is a set of spatio-temporal frames and K is a set of beliefs formed using A and S.

Belief models are assigned semantics based on a modal-logical translation of beliefs into a poly-modal logic that is formed as a fusion of $\mathrm{KD45}_A^{\mathsf{C}}$ (doxastic logic with a common belief operator [4]) for epistemic statuses, $\mathrm{K4}_n$ for subsumption-based spatio-temporal reasoning and $\mathcal{L}_{\mathsf{dom}}$ for content formulas. This gives a straightforward notion of belief model consistency: a belief model is consistent if and only if its modal-logical translation has a model, $\mathbf{B} \models b$ for all beliefs b in $\mathbf{B}$.

The belief model keeps track of the beliefs' evolution in a directed graph called the *history*. The nodes of the history are beliefs and operations on the belief model (such as *retraction*) with (labeled) edges denoting the operations' arguments. The nodes that are beliefs and have no outcoming edges form a consistent, most recent belief model.

B. Attaining common ground

A shared belief of a group G that ϕ implies all private beliefs and all possible attributed beliefs that ϕ within that group. For example, if ϕ is common ground between the human user, h, and robot, r, then (i) implies (ii):

$$\underset{\text{(i)}}{\mathbf{B} \models \mathsf{B}\{\mathsf{r},\mathsf{h}\}/\sigma : \phi} \quad \Rightarrow \quad \underset{\text{(ii)}}{\begin{array}{ll} \mathbf{B} \models \mathsf{B}\{\mathsf{r}\}/\sigma : \phi & \\ \mathbf{B} \models \mathsf{B}\{\mathsf{r}[\mathsf{h}]\}/\sigma : \phi & \\ \mathbf{B} \models \mathsf{B}\{\mathsf{h}\}/\sigma : \phi & * \\ \mathbf{B} \models \mathsf{B}\{\mathsf{h}[\mathsf{r}]\}/\sigma : \phi & * \end{array}}$$

Since (i) and (ii) are inferentially equivalent within belief models, the relation is in fact equivalence. If (ii) holds in the belief model $\mathbf{B}$, it also satisfies (i).

However, the agents' private and attributed beliefs cannot be observed by other agents, as they are not omniscient. The beliefs above marked by asterisk (*) cannot be present in the robot's belief model. The validity of such beliefs can only be *assumed*. An invalidation of the assumptions then invalidates the premise (ii) and thus the conclusion (i). As long as they are not invalidated, agents may act upon them: they may *assume* that common ground has been attained.

But how can these assumptions be in principle mandated or falsified? Given a communication channel C, we consider a class of protocols P_C that supply the means for falsification of the assumptions. If these means are provided, then the protocol is able to reach common ground. We assume that the agents are faithful to Grice's Maxim of Quality [5], i.e. that they are truthful and only say what they believe to be true and for what they have evidence.

C. Abductive inference with assertions

1) Context in abductive inference: Our abductive framework consists of a set of modalised facts $\mathcal{F}$ and a set of rules $\mathcal{R}$. The modal contexts that we use are the following:

- i – *i*nformation. Used to mark the information that is logically true, e.g. description of relational structures.
- e – *e*vent. Used to denote *events* which the robot is trying to understand or produce.
- γ – intention. Marks the intention of an agent's action. In the interpretation phase, it is used to mark the recognised intention. In the generation phase, it is used as a goal in order to find its best possible realisation.
- a – *a*ttentional state. Marks the formulas that are in the attention span. For beliefs, this corresponds to the notion of *foregrounded* beliefs.
- $\mathsf{k}(e)$ – epistemic status. Assigns the predicate an epistemic status (private/attributed/shared).
- $\textsc{During}(\sigma)$ – spatio-temporal frame. Assigns a spatio-temporal frame to the predicate. Together with $[\mathsf{k}(e)]$, the formulas can then be translated into beliefs.

We also include two "technical" contexts that exploit the ability to bring modularity into logic programming following Baldoni et al. [1].

- interpret – understanding phase module.
- generate – generation phase module.

In comparison to S&T's definition of a context [14], we include specific contexts for intentions (γ), epistemic statuses ($\mathsf{k}(e)$) and spatio-temporal frames ($\textsc{During}(\sigma)$), as well as the technical contexts, interpret and generate. While the addition of a context for assigning epistemic statuses and spatio-temporal frames is specific for our purposes and stems from the usage of belief models to model the state of the world and common ground, the addition of the context for distinguishing intentions is more general and allows us to use intentions as an abstract layer.

2) Assertions: We propose a notion of *assertion* for abduction based on *test actions* $\langle F\rangle?$ [2]. Baldoni et al. specify a test as a proof rule. In this rule, a goal F follows from a state $a_1, ..., a_n$ after steps $\langle F\rangle?, p_1, ..., p_m$ if we can establish F on $a_1, ..., a_n$ with answer σ and this (also) holds in the final state resulting fron executing $p_1, ..., p_m$. Using the notion of context as per above, a test $\kappa : \langle F\rangle?$ means we need to be able to verify F in context κ. If we only use axioms A, testing is restricted to observability of facts. An embedded implication $D \supset C$ establishes a *local module*: the clauses D can only be used to prove C. Formulating a test over an embedded implication $\mu : (D \supset \langle C\rangle?)$, we make it explicit that we assume the truth of the statement but require its eventual verification in μ.

Finally, an assertion is the transformation of a test into a partial proof that assumes the verification of the test, while at the same time conditioning the obtainability of the proof goal on the tested statements. Intuitively, $\mu : \langle D\rangle?$ within a proof $\Pi[\langle D\rangle?]$ to a goal C turns into $\Pi[D] \rightarrow C \wedge \mu : D$. Should $\mu : D$ not be verifiable, Π is invalidated.

The verification of an assertion can take various forms. In our system, we check whether a new piece of information can be used to consistently update a belief model (consistency), or to extend a modal model (learning) or weaken it (unlearning).

D. Continual collaborative acting (CCA)

1) The algorithm: Algorithm 1 presents the core of the dialogue management model based on S&T. In the *perception* phase, the agent senses an event e. It tries to understand it in terms of an intention i that results in an update of the belief model from the initial context c_0 to c_1, given the communicative resources r, possible results $Z(c_0)$ to use them in context c_0, and whatever issues are still open to be resolved Σ^π (see below). Given the inferred intention i and potential update to c_1 the agent then tries to carry out this update, as a *verifiable* update.

In *deliberation*, a tacit action based on some private information p may be performed by the agent, giving rise to the context c_3. A public action m is then selected to be performed. In order to communicate the effects of the tacit action, the generation procedure has to use communicative resources $Z(c_2)$ available before the tacit action, while at the same time operating from the context c_3. The result is an intention to act i', which is then realized as $a(i')$.

Our extension of S&T's collaborative acting algorithm [16] uses assertions in abductive inference, to allow for a

```
$\Sigma^\pi = \emptyset$

loop {
Perception
    $e \leftarrow$ SENSE()
    $\langle c_1, i, \Pi \rangle \leftarrow$ UNDERSTAND($r, Z(c_0) \oplus \Sigma^\pi, e$)
    $c_2 \leftarrow$ VERIFIABLE-UPDATE($c_1, i, \Pi$)

Determination and Deliberation
    $c_3 \leftarrow$ ACT-TACITLY($p, c_2$)
    $m \leftarrow$ SELECT($p, c_3$)
    $\langle i', \Pi \rangle \leftarrow$ GENERATE($r, c_3, m, Z(c_2) \oplus \Sigma^\pi$)

Action
    ACT-PUBLICLY($a(i')$)
    $c_4 \leftarrow$ VERIFIABLE-UPDATE($c_3, i', \Pi$)
}
```

Alg. 1. Continual collaborative acting

revision of beliefs once they are falsified. We assume their truth until such a revision occurs. This removes the need for S&T's symmetry assumption. This is represented in the VERIFIABLE-UPDATE operation.

2) Verifiable update: The VERIFIABLE-UPDATE operation operates on the belief model and a structure Σ^π that we call *proof stack*. It is an ordered store of abductive proofs that contain assertions that have not yet been verified or falsified. Given the proof Π, it checks whether there is a proof Π' on the stack whose assertions can be verified using the beliefs of Π. If there are any beliefs in Π' that were falsified, then the Π' should remain on the top: thus, the operation first pushes Π onto the stack and then Π'. The belief model update is then based on those beliefs from Π that have been assumed in the abductive proof and the asserted beliefs beliefs from Π' that have been verified.

VERIFIABLE-UPDATE returns a consistent belief model. Should there be beliefs in the update that cannot be consistently added to the belief model, the operation retracts some beliefs from the belief model so that the model can be updated and stays as descriptive as possible. The retracted beliefs are added to the stack as assertions so that they can be corrected subsequently, or retracted altogether.

3) Grounding using CCA: If the robot (r) understands the human's (h) claim that ϕ in a frame σ, a proof containing the belief $\mathsf{B}\{\mathsf{r}[\mathsf{h}]\}/\sigma : \phi$ is added to the proof stack as an assertion. If the robot can verify ϕ, then this assertion is removed from the stack; the robot can then assume $\mathsf{B}\{\mathsf{h}\}/\sigma : \phi$ per the Maxim of Quality. Similarly, the human's acceptance of the robot's acknowledgment is a verification of an assertion of on the proof stack, on which the robot (again per Maxim of Quality) can assume the belief $\mathsf{B}\{\mathsf{h}[\mathsf{r}]\}/\sigma : \phi$.

Common ground can then be also assumed as long as these beliefs are not contradicted. Should they be contradicted, VERIFIABLE-UPDATE removes them from the belief model, and the assumption of common ground is no longer valid.

IV. Implementation

A. The architecture

The approach has been fully implemented in a cognitive robot architecture. The cognitive architecture integrates sensory and deliberative information-processing components into a single cognitive system, in a modular fashion. The continual collaborative acting (CCA) is implemented as one of these components.

The design of the system is based on the CoSy Architecture Schema (CAS) [7]. CAS is a set of rules that delimit the design of a distributed information-processing architecture in which the basic processing unit is called a *component*. Components related by their function are grouped into *subarchitectures*. Each subarchitecture is assigned a *working memory*, a blackboard from which all the components within the subarchitecture may read or write. Inter-component and inter-subarchitecture communication is achieved by writing to these working memories. The schema is implemented using the CoSy Architecture Schema Toolkit (CAST).

In our scenario, we use a robot in a table-top scenario, observing and manipulating visual objects. The goal is to build a visual categorical models of the objects in the scene. The robot can interact with a human, for example by asking for clarification when it is uncertain about its sensory interpretation of the visual input. This clarification is then used to extend or update the visual models.

The scenario involves the subarchitectures for vision [17], communication ("comsys") and binding [8]. Each subarchitecture's working memory contains specialised representations of the information processed by the associated components. The visual working memory contains regions of interest generated by a segmentor and proto-objects generated by interpreting these regions. The communication subarchitecture working memory contains logical forms generated from parsing utterances. The task of the binding subarchitecture [8] is to combine these subarchitecture-specific data representations into a common a-modal representation. The binding architecture ("binder") uses Bayesian networks to derive a probability distribution over the possible combinations and builds and maintains the belief model in a bottom-up fashion.

B. The abducer

The weighted abduction algorithm as formulated by Stickel [13] and later Baldoni et al. is straightforward to implement within the logic programming paradigm. We have used Mercury, a purely declarative logic/functional programming language that resembles both Prolog and Haskell but that is compiled rather than interpreted [12].

The abducer rule set is currently static and is common for both the understanding and generation phases of the CCA algorithm in which the abducer is used, employing the technical modal contexts to distinguish rules and facts that can only be applied in one of the phases.

C. The CCA component

1) Understanding an observed action: The CCA is implemented as a component within the communication subarchitecture. It is notified of any logical form corresponding to a recognized utterance together with a list of possible bindings of its referential expressions to binder unions appearing on its working memory. This is interpreted as an event observation in the perception phase of the CCA loop. Each of the possible bindings is assigned a probability by the binder. This information is used by the abducer to find the best explanation of the entire utterance.

Currently, the only action that is interpreted as an event by the CCA is a dialogue act by the user. However, the framework can accomodate events recognized by other modalities (such as vision) as well.

2) Clarification requests: If a modality (vision in our scenario) needs to find out more about a certain object from the user, it writes a *clarification request* to the comsys working memory. This is picked up by the CCA, interpreted as a tacit action within the CCA loop. It makes the robot generate a context-aware clarification question. This results in the question core to appear onto the proof stack as an assertion, thus making it a potential belief model update.

3) Verification of asserted beliefs: Modalities can verify the asserted beliefs. For instance, if the user says “the box is blue” (an assertion about the box) the vision subarchitecture is notified of the new assertion appearing on the proof stack and can check whether the information is consistent with its visual model and if not, whether the visual model can be extended or updated. If so, the subarchitecture updates the visual model and notifies the CCA component, which then (as a result of a tacit action) generates an appropriate feedback such as “yes, i can see that”. This change then percolates into the vision working memory and triggers the binder to form an updated belief model.

4) Acting: The public action selection in our implementation is done by using a finite-state automaton that maps recognised communicative intentions to intentions to act. In the future, we would like to employ a POMDP-based action selection [18] rather than a finite-state automaton.

The action is then abductively transformed (GENERATE) to a structure that can be written to a corresponding working memory. Currently, our system only supports communicative actions using the communication subarchitecture.

V. Experimentation

We illustrate our approach on a scenario in which a robot gradually learns more about visual objects that it sees (Figure 1). The interaction is mixed-initiative. Typically the robot drives the dialogue by asking more about what it does not understand. The success of such a dialogue depends strongly on whether the human and the robot can arrive at common ground. This is key in several respects. The robot needs to be able to consistently integrate information it gets through dialogue, into its belief models and visual models. This may concern positive information, resulting in an update of its models, or negative information. In the latter case, the robot needs to revise its belief model, unlearn the incorrect information, and then gather the correct information to learn a better model. Below, we illustrate how the robot can deal with these.

A. Updating beliefs with human information

As the robot observes a new object in the visual scene, it creates a private belief, (1), about this object. The belief is explicitly connected to the a-modal representation u of the object, and is situated in “here and now”, represented by the spatio-temporal frame $\sigma_{\mathsf{here\text{-}now}}$, omitted from the formulas below for the sake of clarity.

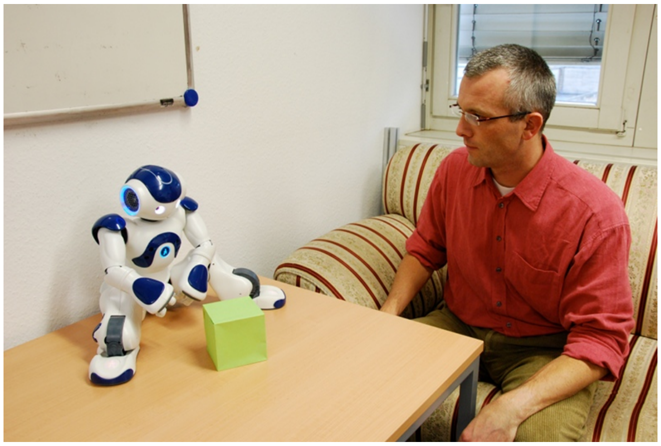

Fig. 1. The setting of the table-top scenario

$$\mathsf{B}\{\mathsf{r}\} : @_u\mathbf{object} \quad (1)$$

After the human has placed the object, he indicates what it is: “This is a box.” The robot creates a semantic representation of this utterance. It uses this information to create a belief it attributes to the human (2): The robot believes the human believes this is a box. This belief is also connected to the visual object, and thus to the robot’s private belief.

$$\begin{array}{rll} \mathsf{B}\{\mathsf{r}\} : & @_u\mathbf{object} & (1) \\ \mathsf{B}\{\mathsf{r}[\mathsf{h}]\} : & @_u\langle \mathit{Type}\rangle\mathbf{box} & \text{(2), assertion} \end{array}$$

The robot can use the type-information to consistently update its visual models. The vision subarchitecture thereby positively verifies the information, represented by a private belief (3) in the belief model.

$$\begin{array}{rll} \mathsf{B}\{\mathsf{r}\} : & @_u\mathbf{object} & (1) \\ \mathsf{B}\{\mathsf{r}[\mathsf{h}]\} : & @_u\langle \mathit{Type}\rangle\mathbf{box} & (2) \\ \mathsf{B}\{\mathsf{r}\} : & @_u\langle \mathit{Type}\rangle\mathbf{box} & (3) \end{array}$$

If the robot then notifies the human of this verification, it can lift the attributed belief (2) with the private belief (3) to a shared belief (4), assuming the information to be grounded.

$$\begin{array}{rll} \mathsf{B}\{\mathsf{r}\} : & @_u\mathbf{object} & (1) \\ \mathsf{B}\{\mathsf{r},\mathsf{h}\} : & @_u\langle \mathit{Type}\rangle\mathbf{box} & (4) \end{array}$$

The robot infers that a box typically has a color – but it does not know what color the box is. Vision accordingly poses an information request to the architecture, which dialogue can help resolve. The request is based on a private belief of the form $\mathsf{B}\{\mathsf{r}\} : @_u\langle \mathit{Color}\rangle\mathbf{unknown}$. Stating color as an assertion means the robot needs information from the human to “verify” it, i.e. fill the gap.

$$\begin{array}{rll} \mathsf{B}\{\mathsf{r}\} : & @_u\mathbf{object} & (1) \\ \mathsf{B}\{\mathsf{r},\mathsf{h}\} : & @_u\langle \mathit{Type}\rangle\mathbf{box} & (4) \\ \mathsf{B}\{\mathsf{r}\} : & @_u\langle \mathit{Color}\rangle\mathbf{unknown} & \text{(5), assertion} \end{array}$$

The human responds cooperatively, saying “It is green.” Abduction yields a proof that this information in principle could answer the question the robot just raised [9]. This gives rise to an attributed belief, with the color information: $\mathsf{B}\{\mathsf{r}[\mathsf{h}]\} : @_u\langle \mathit{Color}\rangle\mathbf{green}$.

$$\begin{array}{rll}
\mathsf{B}\{\mathsf{r}\} : & @_u\mathbf{object} & (1) \\
\mathsf{B}\{\mathsf{r},\mathsf{h}\} : & @_u\langle \mathit{Type}\rangle\mathbf{box} & (4) \\
\mathsf{B}\{\mathsf{r}\} : & @_u\langle \mathit{Color}\rangle\mathbf{unknown} & (5)\text{, assertion} \\
\mathsf{B}\{\mathsf{r}[\mathsf{h}]\} : & @_u\langle \mathit{Color}\rangle\mathbf{green} & (6)\text{, assertion}
\end{array}$$

If vision can now use the information in the updated belief to consistently extend its models, it verifies the assertion. The belief attains shared status.

$$\begin{array}{rll}
\mathsf{B}\{\mathsf{r}\} : & @_u\mathbf{object} & (1) \\
\mathsf{B}\{\mathsf{r},\mathsf{h}\} : & @_u\langle \mathit{Type}\rangle\mathbf{box} & (4) \\
\mathsf{B}\{\mathsf{r},\mathsf{h}\} : & @_u\langle \mathit{Color}\rangle\mathbf{green} & (7)
\end{array}$$

B. Revising the belief model

Now, assume that instead of not knowing the color at all, the robot hypothesizes that the box is yellow. In this case, it asks "Is the box yellow?" based on the belief $\mathsf{B}\{\mathsf{r}\} : @_u\langle \mathit{Color}\rangle\mathbf{yellow}$. If the human now replies with "No, it is not yellow," the robot first creates a corresponding negative belief, and unlearns the classification from its visual models. The negative belief is shared. Next up, it still wants to know what color the box has. The belief model then contains both the shared negative belief (8) and the open private belief about the now unknown color (9).

$$\begin{array}{rll}
\mathsf{B}\{\mathsf{r}\} : & @_u\mathbf{object} & (1) \\
\mathsf{B}\{\mathsf{r},\mathsf{h}\} : & @_u\langle \mathit{Type}\rangle\mathbf{box} & (4) \\
\mathsf{B}\{\mathsf{r},\mathsf{h}\} : & @_u\langle \mathit{Color}\rangle\mathrm{not}(\mathbf{yellow}) & (8) \\
\mathsf{B}\{\mathsf{r}\} : & @_u\langle \mathit{Color}\rangle\mathbf{unknown} & (9)\text{, assertion}
\end{array}$$

The dialogue now returns to a flow similar to the above. If the human responds with "It is green," the robot can again update its belief model and visual models. The robot now holds both a negative shared belief about color ($\mathrm{not}(\mathbf{yellow})$) and a positive shared belief about it ($\mathbf{green}$).

$$\begin{array}{rll}
\mathsf{B}\{\mathsf{r}\} : & @_u\mathbf{object} & (1) \\
\mathsf{B}\{\mathsf{r},\mathsf{h}\} : & @_u\langle \mathit{Type}\rangle\mathbf{box} & (4) \\
\mathsf{B}\{\mathsf{r},\mathsf{h}\} : & @_u\langle \mathit{Color}\rangle\mathrm{not}(\mathbf{yellow}) & (8) \\
\mathsf{B}\{\mathsf{r},\mathsf{h}\} : & @_u\langle \mathit{Color}\rangle\mathbf{green} & (10)
\end{array}$$

All of these beliefs are connected, being anchored to the visual referent we have been talking about. This connection provides a belief history. The robot not only has its current beliefs, it can also introspect how it got there. If the human would now ask, for example to test, whether the robot still thinks whether the object is yellow, the robot can reply "No. It is green." This makes fully transparent the chain of shared beliefs that the robot has, pertaining to the box object.

VI. Conclusions

We presented an approach to processing situated dialogue in human-robot interaction set in a larger collaborative activity. The approach both looks at what utterances are about, and why they are or should be uttered: intension and intention are put on par. The approach uses weighted abduction to drive processing. This allows for a smooth integration with probabilistic interpretation hypotheses that we get from other forms of processing, e.g. binding or vision.

Currently, we are investigating how we can combine this approach with plan- and intention recognition to achieve a close integration with collaborative action planning, and with POMDP-based action selection. The latter would help us to select actions even when interpretation does not yield enough information to completely interpret an utterance.

VII. Acknowledgments

The research leading to these results has received funding from the European Union's 7th Framework Programme [FP7/2007-2013] under grant agreement No. 215181 (CogX).

References

[1] M. Baldoni, L. Giordano, and A. Martelli. A modal extension of logic programming: Modularity, beliefs and hypothetical reasoning. *Journal of Logic and Computation*, 8(5):597–635, 1998.

[2] M. Baldoni, L. Giordano, A. Martelli, and V. Patti. A modal programming language for representing complex actions. In A. Bonner, B. Freitag, and L. Giordano, editors, *Proceedings of the 1998 JICSLP'98 Post-Conference Workshop on Transactions and Change in Logic Databases (DYNAMICS'98)*, pages 1–15, 1998.

[3] M. Brenner and B. Nebel. Continual planning and acting in dynamic multiagent environments. *Journal of Autonomous Agents and Multiagent Systems*, 2008.

[4] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge*. MIT Press, 1995.

[5] H.P. Grice. Logic and conversation. *Syntax and Semantics*, 3:41–58, 1975.

[6] B.J. Grosz and C.L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[7] N. Hawes and J. Wyatt. Engineering intelligent information-processing systems with CAST. *Advanced Engineering Infomatics*, 24(1):27–39, 2010.

[8] H. Jacobsson, N.A. Hawes, G.J.M. Kruijff, and J. Wyatt. Crossmodal content binding in information-processing architectures. In *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Amsterdam, The Netherlands, March 12–15 2008.

[9] G.J.M. Kruijff and M. Brenner. Phrasing questions. In *Proceedings of the AAAI 2009 Spring Symposium on Agents That Learn From Humans*, 2009.

[10] K. Lochbaum, B.J. Grosz, and C.L. Sidner. Discourse structure and intention recognition. In R. Dale, H. Moisl, , and H. Somers, editors, *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*. Marcel Dekker, New York, 1999.

[11] M.J. Pickering and S. Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–225, 2004.

[12] Z. Somogyi, F. Henderson, and T. Conway. Mercury: An efficient purely declarative logic programming language. In *Proceedings of the Australian Computer Science Conference*, pages 499–512, Feb 1995.

[13] M.E. Stickel. A prolog-like inference system for computing minimum-cost abductive explanations in natural-language interpretation. Technical Report 451, AI Center, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, Sep 1988.

[14] M. Stone and R.H. Thomason. Context in abductive interpretation. In *Proceedings of EDILOG 2002: 6th workshop on the semantics and pragmatics of dialogue*, 2002.

[15] M. Stone and R.H. Thomason. Coordinating understanding and generation in an abductive approach to interpretation. In *Proceedings of DIABRUCK 2003: 7th workshop on the semantics and pragmatics of dialogue*, 2003.

[16] R.H. Thomason, M. Stone, and D. DeVault. Enlightened update: A computational architecture for presupposition and other pragmatic phenomena. In D. Byron, C. Roberts, and S. Schwenter, editors, *Presupposition Accommodation*. to appear.

[17] A. Vrečko, D. Skočaj, N. Hawes, and A. Leonardis. A computer vision integration model for a multi-modal cognitive system. In *IEEE/RSJ International Conference on Intelligent RObots and Systems*, pages 3140–3147, 2009.

[18] S. Young, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. The hidden information state model: a practical framework for POMDP-based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174, 2009.