HAL
open science

Dialogue in situated environments: A symbolic approach to perspective-aware grounding, clarification and reasoning for robot

Séverin Lemaignan, Raquel Ros, Rachid Alami

To cite this version:

Séverin Lemaignan, Raquel Ros, Rachid Alami. Dialogue in situated environments: A symbolic approach to perspective-aware grounding, clarification and reasoning for robot. Robotics, Science and Systems, Grounding Human-Robot Dialog for Spatial Tasks workshop, Jul 2011, United States. pp.1. hal-00667030

HAL Id: hal-00667030

https://hal.science/hal-00667030v1

Submitted on 6 Feb 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Dialogue in situated environments: A symbolic approach to perspective-aware grounding, clarification and reasoning for robots

Séverin Lemaignan, Raquel Ros, Rachid Alami

CNRS - LAAS, Université de Toulouse, UPS, INSA, INP, ISAE, LAAS, F-31077 Toulouse, France
{slemaign, rrosespi, rachid}@laas.fr

***Abstract*—Interacting with a robot in a shared space requires not only a shared model (*i.e.,* environment entities must be described with the same symbols), but also a model of mutual knowledge: both the robot and the human need to figure out what the other knows, sees or can do in the environment to appropriately behave. We present here our efforts to let the robot grounds verbal interaction with a human in a physical, situated environment. We propose a knowledge-oriented architecture, where perceptions from different point of views (from the robot itself, from the human, etc.) are turned into symbolic facts that are stored in different cognitive models and reused in a newly designed module for dialogue grounding.**

I. Grounding human interaction into the robot knowledge

In [1], Baron-Cohen, Leslie and Frith introduced the idea of *theory of mind* to explain cognitive deficit in autistic children: the lack of such theory would prevent them to build a mental model of what their interactors perceive and think, leading to communication issues and strong social impairment.

Figure 1 pictures a situation where such an ability to model different *perspectives* is needed for a successful interaction: the cardboard box on the table contains one video tape visible to the robot, but not to the human, while another one is on the table visible to both agents. In this context, when the human says " – Jido, give me the tape", he is obviously referring to the tape that he sees, on the table.

For the robot, this kind of reasoning can only be made if the robot is able to build and maintain an independent cognitive model for the human, and use it to resolve the concept involved in the verbal interaction.

Our work focuses on these issues: what are the prerequisites for such a human sentence — "Jido, give me the tape" — to be understood by the robot, correctly interpreted in the spatial context of the interaction, and ultimately transformed into an action?

This implies first to understand the semantics of the sentence: What does "*Jido*" refers to? What is "*give*"? What is "*me*"? And "*the tape*"? Working in a situated context, we want furthermore to *resolve* these semantics atoms, *i.e.,* ground them in the sensory-motor space of the robot. For instance, in this context "*tape*" refers to an artifact of type `VideoTape` that is known by the human (thus excluding the video tape hidden inside the box).

Fig. 1. Interacting with the robot in an everyday setup: the human asks for help in vague terms, the robot takes into account the human's spatial perspective to refine its understanding of the question.

Extracting the intent of a sentence, resolving its members and turning them into content processable by the robot is a difficult challenge in the general case. We base our approach on three distinct, inter-related cognitive functions:

1) Physical environment modeling and *spatial reasoning* (grouped under the term *situation assessment*, Fig.2) are in charge of building and maintaining a coherent model of the physical world [8]. The geometric model is used to compute several spatial properties of the scene that actually convert the original sensory data into symbolic beliefs. We focus on two types of properties which we believe are essential in Human-Robot interaction: *i*) agent's capabilities, which correspond to actions an agent can perform on objects or other agents (*e.g.* see, reach, point at) and *ii*) object's locations, which correspond to positions of objects with respect to other objects or agents. In both cases the different properties are computed from the point of view of the different agents involved in the scene. This allows the robot to model the world and reason from different perspectives, in a similar way we humans build different mental states for each person we interact with.

This model is realistic in the sense that it relies on accurate 3D models of both manipulable objects and humans. It also has

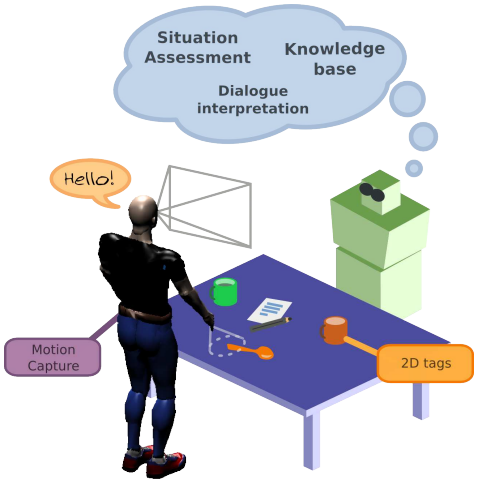

Fig. 2. The robot builds its symbolic model of the world through *situation assessment*. Percepts come from motion capture (to track human posture) and tag-based object recognition. The human field of view, pointing area (grey dots on the table) as well as spatial relations between objects are computed and stored in a ontology-based knowledge base.

a dedicated mechanism to manage disappearing or occluded objects. Thus, if an object is not currently visible from an agent point of view, but the robot is aware that the agent has seen it before, then it is considered as a "known" object, but occluded for that agent. On the contrary, if the agent does not know about the existence of the object, then it is only considered as unknown (and not occluded).

2) Knowledge representation and management: the robot is endowed with an active knowledge base, based on an ontology [5]. It provides a logically sound symbolic model of the robot's beliefs on the world, as well as models for each cognitive agent the robot interacts with. Each of these models is independent and logically consistent. This enables reasoning on different perspectives of the world that would be considered otherwise inconsistent (for instance, an object can be visible for the robot but not for the human. This object can have at the same time the property `isVisible` **`true`** and `isVisible` **`false`**, in two different models). Used in combination with the situation assessment framework, the robot is thus able to maintain different models of the world, one per agent. This proves an essential feature [7, 4] to enable perspective-aware grounding of natural language, as mentioned below.

Our knowledge base relies on OWL ontologies (a decidable subset of the predicate logics) and features continuous storage, querying and event triggering over the pool of facts known by the robot. It also provides an initial pool of general facts, the ORO Commonsense Ontology[1]. This ontology contains a set of concepts, relationships between concepts and rules that defines the "cultural background" of the robot, *i.e.,* the a priori known concepts (currently, the commonsense ontology defines about 100 concepts and over 80 properties that link these concepts in our domain). In this work, this commonsense knowledge is focused on the requirement of human-robot interactions in everyday environments, but also contains generic concepts such as `thing`, `object`, `location` and relationships between those.

3) Dialogue input processing, including natural language parsing capabilities, disambiguation routines and interactive concept anchoring. We focused our efforts on three classes of utterance, commonly found in human-robot interaction: *statements* (*i.e.,* new facts the human wants to inform the robot), *orders* (or more generically *desires*) and *questions on declarative knowledge* (where their answers do not require explicit planning). This would roughly cover the *representative* (sometimes referred as *assertives*) and *directives* type of illocutionary acts, in Searle classification.

A. Contributions

In [7], Roy summarizes what he sees as the main challenges to be tackled for successful language interpretation for robots: cross-modal representation systems, association of words with perceptual and action categories, modeling of context, figuring out the right granularity of models, integrating temporal modeling and planning, the ability to match past (learned) experiences with the current interaction and the ability to take into account the human perspective. We have attempted to tackle some of these challenges (leaving aside temporal modeling and learning) in our approach.

Compared to previous contributions in the field of natural language processing for robots [2, 3, 6], our efforts have two foci: *(1)* integration between language processing and perception of the environment and the humans, from several perspectives; and *(2)* realistic human-robot interactions: open speech; complex, dynamic, partially unknown human environments; fully embodied autonomous robots with manipulation abilities.

Note that we do not claim any contribution to the field of computational linguists itself (see [4] for a survey of formal approaches to natural language processing in the robotics field). Our contribution regards the grounding (we call it *resolution*) of concepts involved in the human discourse through the robot's own knowledge.

II. The Natural Language Grounding Process

Verbal interaction with humans presents two categories of challenges: syntactic ones, and semantic ones. The robot must be able to process and analyze the structure of human utterances, *i.e.,* natural language sentences, and then make sense of them. As stated in the introduction, we process three categories of sentences: *statements*, *desires* and *questions*. In our approach, the grounding of the human discourse consists in extracting either the *informational* content of the sentence to produce statements or its *intentional* content (*i.e.,* performative value) to collect orders and questions.

We have developed a dedicated module called Dialogs[2] that processes human input in natural language, grounds the

[1]The knowledge base and this ontology are open-source projects. http://oro.openrobots.org.

[2]http://dialogs.openrobots.org

concepts in the robot's knowledge and eventually translates the discourse in a set of declarative OWL/RDF statements.

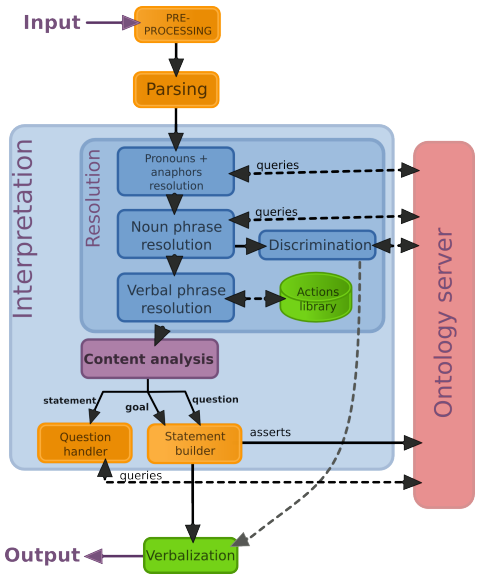

Fig. 3. The DIALOGS module has three main steps: the parsing, the interpretation and the verbalization. The interpretation module is responsible for both the *resolution* and the semantic content *analysis and translation*.

As shown in Figure 3, the DIALOGS module is composed of three main blocks. The user's input is first pre-processed and then parsed. The parser is a custom-made, rule-based (*i.e.,* grammar-free) tool that extracts the grammatical structure from the user's sentence.

The result of the parsing is then sent to the *interpretation* module, the core of the tool. Interpretation consists in three distinct operations: the sentence *resolution* (concepts grounding), the *content analysis* (what is the intent of the utterance: information, question or desire) and the *statement building* (translation into RDF statements).

Interpretation tightly relies on the communication with the knowledge base: all the concepts the robot manipulates are stored in the *ontology server* and retrieved through logical queries, except for the verbs that are currently stored in a dedicated library (the *action library* in the diagram).

The last block allows for natural recomposition of sentences, before verbalization (either textual or via a text-to-speech interface).

III. CASE STUDY

The following case study illustrates the approach: Tom and Jerry are moving to London, so they are packing things in boxes. The scenario takes places in the living-room, where Jido (our robot) is observing while they move things here and there. To assess the reasoning abilities of the robot they ask Jido for information (entered through keyboard). Ideally, the robot should also perform actions when required (*e.g.* hand an object when asking "give me..."). However, since it is out of the scope of this work, we do not include any motion from the robot's side.

```
>> IMPERATIVE
Verbal Group: give (present simple)
        direct objects:
          Nominal Group: the videotape

        indirect objects:
          Nominal Group: me
```

Fig. 4. Raw output of the DIALOGS parser for the sentence "Give me the videotape", before grounding of groups.

Perception of objects is done through a tag-based system[3] and humans are detected through motion capture. The robot knowledge base is pre-loaded with the *ORO Commonsense Ontology*. We next describe in detail two situations where we can follow the internal robot's reasoning and the interaction with the users.

1) Implicit disambiguation through visual perspective taking: Tom enters the room while carrying a big box (Figure 1, page 1). He approaches the table and asks Jido to handle him the video tape: "Jido, can you give me the video tape". After parsing (Figure 4), the DIALOGS module queries the ontology to identify the object the human is referring to: `?obj` **`type`** `VideoTape`.

There are two video tapes in the scene: one on the table, and another one inside the cardboard box. Thus, the knowledge base returns both: ⇒ `?obj = [videoTape1, videoTape2]`.

However, only one is visible for Tom (the one on the table). Thus, although there is an ambiguity from the robot's perspective (since it can see both video tapes), based on the perspective of its human partner it infers that Tom is referring to the video tape on the table, and not the one inside the box which is not visible from his view. Therefore, non-visible objects are removed obtaining: `?obj =[videoTape1]`.

Since only one object is available, the robot infers that the human refers to it and would eventually execute the command, *i.e.,* give it to the human. Alternatively, the robot could first verify with the human if that was the object being referred to or not before proceeding to execute the action.

Fig. 5. Jerry asks Jido for the content of the box by pointing at it.

2) Explicit disambiguation through verbal interaction and gestures: Figure 5 depicts a situation where Jerry enters the

[3]Each tag is associated to an unique id, without any specific semantic attached to it. Semantics – like the type of the object – are either stored in a scenario specific ontology loaded at startup or teached online to the robot.

living room without knowing where Tom had placed the video tapes. So he first asks Jido: "What's in the box?". Before the robot can answer the question it has to figure out which box Jerry is talking about. Similar to the previous situation, there are two available boxes:

```
?obj type box
    ⇒ ?obj = [cardBoardBox, toolbox]
```

However both are visible and the cognitive ambiguity resolution cannot be applied. The only option is to ask Jerry which box he is referring to: "Which box, the toolbox or the cardboard box?" Jerry could now simply answer the question. Instead, he decides to point at it while indicating: "This box". The robot's perception identifies the `cardBoardBox` as being pointed at and looked at by the human and updates the ontology with this new information using a rule available in the commonsense ontology (**pointsAt**(?ag, ?obj) ∧ **looksAt**(?ag, ?obj) → **focusesOn**(?ag, ?obj)) The DIALOGS module is then able to merge both sources of information, verbal ("this") and gestural to distinguish the box Jerry refers to.

Finally, DIALOGS queries the ontology about the content of the box and the question can be answered: "Jido-E".

```
?obj isIn cardBoardBox
    ⇒ ?obj = videoTape2
```

Note that the object's label is used instead of its ID. This way we enhance interaction using familiar names given by the users.

At this point Jerry wants to know where the other tape is, and that is exactly what he asks Jido: "And where is the other tape?". In this occasion, the DIALOGS module is able to interpret that Jerry is not referring to the video which they were just talking about, but to the other one:

```
?obj type VideoTape
?obj differentFrom videoTape2
    ⇒ ?obj = [videoTape1]
```

Since there is only one possible "other" video (there are only two videos in the scene), it can directly answer Jerry: "The other tape is on the table and next to the toolbox."

```
videoTape1 isOn table
videoTape1 isNextTo toolbox
```

IV. CONCLUSION

We propose the DIALOGS module that converts natural language utterances into either symbolic facts (OWL statements) or natural language answers, depending on the intent conveyed by the original sentence.

Grounding of referent is done by relying on a situation assessment reasoner and a symbolic knowledge base that are able to compute and store several *perspectives* of the world state, one for each agent. Our system takes also into account non-verbal communication cues, like gaze or pointing gestures.

Using so-called thematic roles and the symbolic reasoning capabilities of the knowledge base, semantic correctness of utterances can be checked by the robot who can react accordingly.

However, when looking back to the proposed agenda (Section I-A), much remains to be done.

For instance, the natural language parsing abilities of the system are largely *ad-hoc*, and were not designed to scale beyond the domain of simple everyday dialogue for joint interaction (extensions are however planned, like automatic fetching of thematic roles from online libraries like VERBNET or support for parsing new rules expressed in natural language and adding them to the symbolic model, like "If someone looks at something then he sees it.").

We also need to enrich the *cross-modalities* by better taking into account the physical behaviour of the speaker and its evolution in time. New sensors like the Microsoft Kinect open interesting opportunities in this domain.

Finally, we need to tackle as well the issue of rigorous repeatable evaluation of the framework, validated by user-studies. Due to the intricate combination between natural language and dynamic, semantic-rich physical environments, a good methodology for evaluation still has to be designed.

ACKNOWLEDGMENTS

Part of this work has been conducted within the EU CHRIS project (`http://www.chrisfp7.eu/`) funded by the E.C. Division FP7-IST under Contract 215805.

REFERENCES

[1] S. Baron-Cohen, A.M. Leslie, and U. Frith. Does the autistic child have a "theory of mind" ? *Cognition*, 1985.

[2] T. Brick and M. Scheutz. Incremental natural language processing for HRI. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, 2007.

[3] S. Huwel, B. Wrede, and G. Sagerer. Robust speech understanding for multi-modal human-robot communication. In *The 15th IEEE International Symposium on Robot and Human Interactive Communication*, 2006.

[4] G.J.M. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, H. Zender, I. Kruijff-Korbayová, and N. Hawes. Situated dialogue processing for human-robot interaction. *Cognitive Systems*, pages 311–364, 2010.

[5] S. Lemaignan, R. Ros, L. Mösenlechner, R. Alami, and M. Beetz. ORO, a knowledge management platform for cognitive architectures in robotics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

[6] N. Mavridis and D. Roy. Grounded situation models for robots: Bridging language, perception, and action. In *AAAI-05 Workshop on Modular Construction of Human-Like Intelligence*, 2005.

[7] D. Roy and E. Reiter. Connecting language to the world. *Artificial Intelligence*, 2005.

[8] E.A. Sisbot, R. Ros, and R. Alami. Situation assessment for human-robot interaction. In *20th IEEE International Symposium in Robot and Human Interactive Communication*, 2011.