Users' Belief Awareness in Reinforcement Learning-based Situated Human-Robot Dialogue Management

Emmanuel Ferreira, Grégoire Milliez, Fabrice Lefèvre, Rachid Alami

Users' Belief Awareness in Reinforcement Learning-based Situated Human-Robot Dialogue Management

Emmanuel Ferreira, Grégoire Milliez, Fabrice Lefèvre and Rachid Alami

Abstract Others can have a different perception of the world than ours. Understanding this divergence is an ability, known as perspective taking in developmental psychology, that humans exploit in daily social interactions. A recent trend in robotics aims at endowing robots with similar mental mechanisms. The goal then is to enable them to naturally and efficiently plan tasks and communicate about them. In this paper we address this challenge extending a state-of-the-art goal-oriented dialogue management framework, the Hidden Information State (HIS). The new version makes use of the robot's awareness of the users' belief in a reinforcement learning-based situated dialogue management optimisation procedure. Thus the proposed solution enables the system to cope with the communication ambiguities due to noisy channel but also with the possible misunderstandings due to some divergence among the beliefs of the robot and its interlocutor in a Human-Robot Interaction (HRI) context. We show the relevance of the approach by comparing different handcrafted and learnt dialogue policies with and without divergent belief reasoning in an in-house Pick-Place-Carry scenario by mean of user trials in a simulated 3D environment.

1 Introduction

When robots and humans share a common environment, previous works have shown how much enhancing the robot's perspective taking and intention detection abilities improves its understanding of the situation, and leads to more appropriate and efficient task planning and interaction strategies [2, 3, 13]. As part of the theory of mind, perspective taking is a widely studied ability in developmental literature.

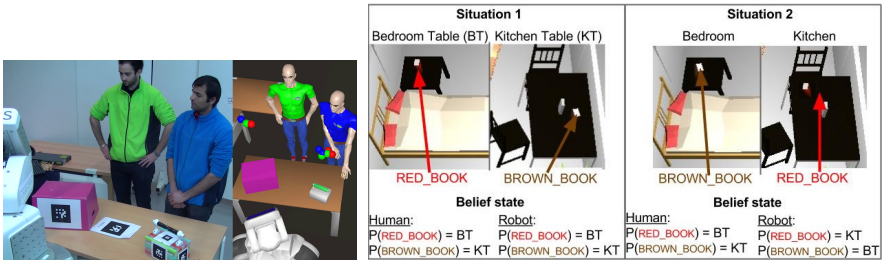
Ferreira Emmanuel and Fabrice Lefèvre
LIA - University of Avignon, France e-mail: {emmanuel.ferreira,fabrice.lefevre}@univ-avignon.fr

Grégoire Milliez and Rachid Alami
CNRS LAAS - University of Toulouse e-mail: {gregoire.milliez,rachid.alami}@laas.fr

This broad term encompasses 1) perceptual perspective taking, whereby human can understand that other people see the world differently, and 2) conceptual perspective taking, whereby humans can go further and attribute thoughts and feelings to other people [1]. Tversky and al. [19] explain to what extent switching between perspectives rather than staying in an egocentric position can improve the overall dialogue efficiency in a situated context. Therefore, to make robots more socially competent, some research aims to endow robots with this ability. Among others, Breazeal et al. [2] present a learning algorithm that takes into account information about a teacher's visual perspective in order to learn specific coloured buttons activation/deactivation patterns, and Trafton et al. [18] use both visual and spatial perspective taking to find out the referent indicated by a human partner. In the present study, we specifically focus on a false belief task as part of the conceptual perspective taking. Formulated in [20], this kind of task requires the ability to recognize that others can have beliefs about the world that differ from the observable reality. Breazal et al. [3] proposed one of the first human-robot implementation and proposed some more advanced goal recognition skills relying on this false belief detection. In [13], a Spatial Reasoning and Knowledge component (SPARK) is presented to manage separate models for agent belief state and used to pass the Sally and Anne test [1] on a robotic platform. This test is a standard instance of false belief task where an agent has to guess the belief state of an other agent with a divergent belief mind state. The divergence in this case arises from modifications of the environment which one agent is unaware of and which are not directly observable, for instance displacement of objects hidden to this agent (behind another object for instance).

Considering this, to favour the human intention understanding and improve the overall dialogue strategy, we take benefit of the divergent belief management into the multimodal situated dialogue management problem. To do so, we rely on the Partially Observable Markov Decision Process (POMDP) framework. This latter is becoming a reference in the Spoken Dialogue System (SDS) field [21, 17, 14] as well as in HRI context [15, 11, 12], due to its capacity to explicitly handle parts of the inherent uncertainty of the information which the system (the robot) has to deal with (erroneous speech recognizer, falsely recognised gestures, etc.). In the POMDP setup, the agent maintains a distribution over possible dialogue states, the belief state, all along the dialogue course and interacts with its perceived environment using a Reinforcement Learning (RL) algorithm so as to maximise some expected cumulative discounted reward [16]. So our goal here is to introduce the divergence notion into the belief state tracking and add some means to deal with it in the control part.

The remainder of the paper is organised as follows. Section 2 gives some details about how an agent knowledge model can be maintained in a robotic system; in Section 3 our extension of a state-of-art goal-oriented POMDP dialogue management framework, the Hidden Information State (HIS), is presented to take into account users' beliefs state; in Section 4 the proposed Pick-Place-Carry false belief scenario used to exemplify the benefit of both taking account of the perspective taking ability and its integration in a machine learning scheme is introduced. In the same section, the current system architecture and the experimental setup employed are given. The



(a) (b)

Fig. 1: (a) Real users in front of the robot (right) and the virtual representation built by the system (left). (b) Divergent belief example with belief state.

user trial results obtained with a learnt and an handcrafted belief-aware system are compared in Section 5 with systems lacking perspective taking ability. Finally, in Section 6 we discuss some conclusions and give some perspectives.

2 Agent knowledge management

As mentioned in the introduction, the spatial reasoning framework SPARK is used for situation assessment and spatial reasoning. We will briefly recap here how it works, for further details please refer to [13]. In our system, the robot collects data about three different entities to virtually model its environment: objects, humans and *proprioceptions* (its own position, posture, etc.). Concerning objects, a model of the environment is loaded at startup to obtain the positions of static objects (e.g. walls, furnitures, etc.). Other objects (e.g. mug, tape, etc.) are considered as movable. Their positions are gathered using the robot's stereo vision. Posture sensors, such as Kinect, are used to obtain the position of humans. These perception data allow the system to use the generated virtual model for further spatial-temporal reasoning. As an example, the system can reason on why an object is not perceived any more by a participant and decide to keep its last known position if it recognizes a situation of occlusion, or remove the object from its model if there is none.

Figure 1 (a) shows a field experiment with the virtual environment built by the system from the perception data collected and enriched by the spatial reasoner. The latter component is also used to generate facts about the objects relative position and agents' affordances. The relative position such as *isIn*, *isNextTo*, *isOn* are used for multimodal dialogue management as a way to solve referents in users' utterances, but also for a more natural dialogue description of the objects position in the robot's responses. Agents' affordances come from their ability to perceive and reach objects. The robot is calculating its own capability of perception according to the actual data it gets from the object position and recognition modules. For reachability, the robot computes if it is able to reach the object with its grasping joints.

To compute the human's affordances the robot applies its perspective taking ability. In other words, the robot has to estimate what is visible and reachable for the human according to her current position. For visibility, it computes which objects are present in a cone, emerging from human's head. If the object can be directly linked to the human's head with no obstacle and if it is in the field of the view cone, then it is assumed that the human sees the object and hence has knowledge of its true position. If an obstacle is occluding the object, then it won't be visible for the human. Concerning the reachability, a threshold of one meter is used to determine if the human can reach an object or not.

The facts generation feature allows the robot to get the information about the environment, its own affordances, and the human's affordances. In daily life, humans get the information about the environment through perception and dialogue. Using the perspective taking abilities of our robot, we can compute a model of each human's belief state according to what she perceived or what the robot has told her about the environment. Then two different models of the world are considered: one for the world state from the robot perception and reasoning and one for each human's belief state (computed by the robot according to what the human perceived). Each of these models is independent and logically consistent. In some cases, the robot and the human models of the environment can diverge. As an example, if an object O has a property P with a value A, if P's value changed to B and the human had no way to perceive it when it occurred, the robot will have the value B in its model ($P(O) = B$) while the human will still have the value A for the property P ($P(O) = A$). This value shouldn't be updated in the human model until the human is actually able to perceive this change or until the robot informs him. In our scenario, this reasoning is applied to the position property.

We introduce here an example of false belief situation (fig. 1 (b)). A human sees a red book (RED_BOOK) on the bedside table BT. She will then have this property in his belief state: P(RED_BOOK) = BT. Now, while this human is away (has no perception of BT), the book is swapped with an other brown one (BROWN_BOOK) from the kitchen table KT. In this example, the robot explores the environment and is aware of the new position values. The human will keep this belief until she gets a new information on the current position of RED_BOOK. This could come from actually seeing RED_BOOK on the position KT or seeing that RED_BOOK is not any more in BT (in which case the position property value will be updated to an *unknown* value). Another way to update this value is for the robot to explicitly inform the user of the new position.

In our system we mainly focused on position properties but this reasoning could be straightforwardly extended to other properties such as who manipulated an object, its content, temperature, etc. Obviously if this setup generalises quite easily to false beliefs about individual properties of elements of the world, more complex divergence configurations that might arise in daily interactions, for instance due to prior individual knowledge, still remain out of range and should be addressed by future complementary works.

3 Belief Aware Multimodal Dialogue Management

As mentioned earlier, an important aspect of the approach is to base our user belief state management on the POMDP framework [9]. It is a generalisation of the fully-observable Markov Decision Process (MDP), that was first employed to determine an optimal mapping between situations (dialogue states) and actions for the dialogue management problem in [10]. We try hereafter to recall some of the principles of this approach pertaining to the modifications that will be introduced. More comprehensive descriptions should be sought in the cited papers. This framework maintains a probability distribution over dialogue states, called belief states, assuming the true one is unobservable. By doing so, it explicitly handles parts of the inherent uncertainty on the information conveyed inside the Dialogue Manager (DM) (e.g. error prone speech recognition and understanding processes). Thus, POMDP can be cast as a continuous space MDP. The latter is a tuple $< B, A, T, R, \gamma >$, where B is the belief state space (continuous), A is the discrete action space, T is a set of Markovian transition probabilities, R is the immediate reward function, $R : B \times A \times B \rightarrow \Re$ and $\gamma \in [0,1]$ the discount factor (discounting long term rewards). The environment evolves at each time step t to a belief state b_t and the agent picks an action a_t according to a policy mapping belief states to actions, $\pi : B \rightarrow A$. Then the belief state changes to b_{t+1} according to the Markovian transition probability $b_{t+1} \sim T(.|b_t, a_t)$ and, following this, the agent received a reward $r_t = R(b_t, a_t, b_{t+1})$ from the environment. The overall problem of this continuous MDP is to derive an optimal policy maximising the reward expectation. Typically the averaged discounted sum over a potentially infinite horizon is used, $\sum_{t=0}^{\infty} \gamma^t r_t$. Thus, for a given policy and start belief state b, this quantity is called the value function: $V^{\pi}(b) = E[\sum_{t \geq 0} \gamma^t r_t | b_0 = b, \pi] \in \Re^B$. V^* corresponds to the value function of any optimal policy π^*. The Q-function may be defined as an alternative to the value function. It adds a degree of freedom on the first selected action, $Q^{\pi}(b,a) = E[\sum_{t \geq 0} \gamma^t r_t | b_0 = b, a_0 = a, \pi] \in \Re^{B \times A}$. As well as V^*, Q^* corresponds to the action-value function of any optimal policy π^*. If it is known, an optimal policy can be directly computed by being greedy according to Q^* , $\pi^*(b) = \arg\max_a Q^*(b,a) \forall b \in B$.

However, real-world POMDP problems are often intractable due to their dimensionality (large belief state and action spaces). Among other techniques, the HIS model [21] circumvents this scaling problem for dialogue management by the use of two main principles. First, it factors the dialogue state into three components: the user goal, the dialogue history and the last user act (see Figure 2). The possible user goals are then grouped together into *partitions* on the assumption that all goals from the same partition are equally probable. These partitions are built using the dependencies defined in a domain-specific ontology and the information extracted all along the dialogue from both the user and the system communicative acts. In the standard HIS model, each partition is linked to matching database entities based on its static and dynamic properties that corresponds to the current state of the world (e.g. colour of an object vs spatial relations like *isOn*). The combination of a partition, the associated dialogue history, which corresponds here to a finite state

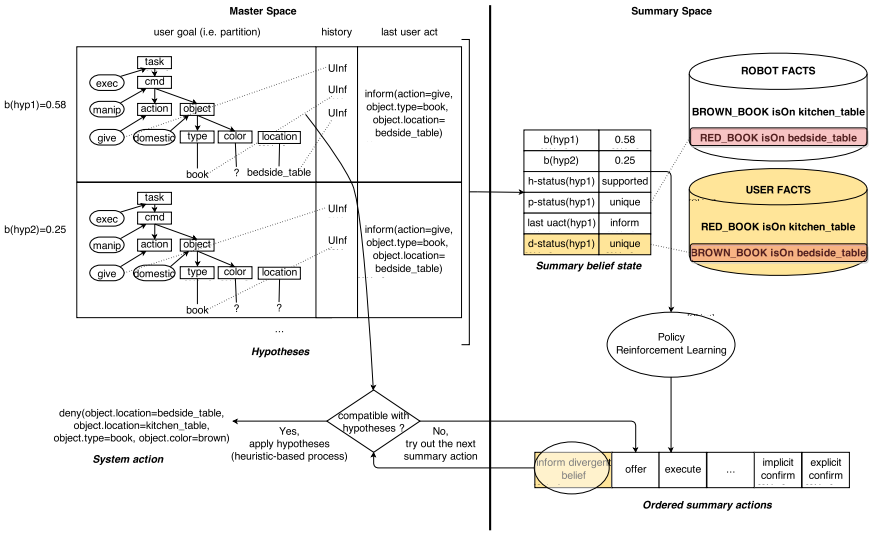


Fig. 2: Overview of the HIS extension to take into account divergent belief.

machine that keeps track of the grounding status for each convoyed piece of information (e.g. informed or grounded by the user), and a possible last user action forms a dialogue state hypothesis. A probability distribution $b(hyp)$ over the most likely hypotheses is maintained during the dialogue and this distribution constitutes the POMDP's belief state. Second, HIS maps both the belief space (hypotheses) and the action space into a much reduced summary space where RL algorithms are tractable. The summary state space is the compound of two continuous and three discrete values. Continuous values are the probabilities of the two-first hypotheses $b(hyp1)$ and $b(hyp2)$ while the discrete ones, extracted from the top hypothesis, are the type of the last user act (noted *last uact*), a partition status (noted *p-status*) database matching status related to the corresponding goal and a history status (noted *h-status*). Likewise system dialogue acts are simplified in a dozen of summary actions like *offer*, *execute*, *explicit-confirm* and *request*. Once the summary actions are ordered by their $Q(b,a)$ scores in descending order by the policy, an handcrafted process checks if the best scored action is compatible with the current set of hypotheses (e.g. for the *confirm* summary act this compatibility test consists in checking if there is something to confirm in the top hypothesis). If they are compatible, an heuristic-based method maps this action back to the master space as the next system response. If not, the process is pursued using the next best scored summary action until a possible action is found.

The standard HIS framework can properly handle misunderstandings due to noise in the communicative channel. However, misunderstandings can also be introduced in cases where the user has false beliefs, impacting negatively her communicative acts. HIS has no dedicated mechanism to deal with such a situation and so it should react as in front of a classical uncertainty by asking the user to confirm hypotheses until the request can match the reality, although it could have be resolved since the

first turn. Therefore having an appropriate mechanism should improve the quality and efficiency of the dialogue, preventing user to pursue her goal with an erroneous statement.

So, as illustrated in Figure 2 and highlighted with the orange items, we propose to extend the summary belief state with an additional status, the *divergent belief* status (noted *d-status*), and an additional summary action, *inform divergent belief*. The *d-status* is employed to trigger the presence of false belief situations by matching the top partition with user facts compiled by the system (see Sec. 2) and as such trying to highlight some divergences between the user and the robot points of view. Both the user and the robot facts (from the belief models, not to be mistaken with the belief state related to the dialogue representation) are considered as part of the dynamic knowledge resource and are maintained independently of the internal state of the system with the techniques described in Sec. 2. Here we can observe in Figure 2 that the top partition is about a book located on the bedside table. In the robot model of the world (i.e. robot facts) this book is identified as a unique entity, RED_BOOK, and *p-status* is set to *unique* accordingly. However, in the user model it is identified as BROWN_BOOK. This situation can be considered as divergent and *p-status* is set to *unique* too because there is one possible object that corresponds to that description in the user model. In this preliminary study *d-status* can only be *unique* or *non-unique*. Further studies may consider more complex cases. The new summary action is employed for appropriate resolution and removal of the divergence. The (real) communicative acts associated to this (generic) action relies on expert design. In this first version, if this action is compatible with the current hypotheses and thus picked up by the system, it explicitly informs the user of the presence and the nature of the divergence. To do so, the system uses a *deny* dialogue act to inform the user about the existence of a divergent point of view and let the user agree on the updated information. Consequently, the user may pursue its original goal with the correct property instead of the obsolete one. This process is also illustrated in Figure 2 when the *inform divergent belief* action is mapped back to the master space.

4 Scenario & Experimental Setup

In order to illustrate the robot's ability to deal with user's perspective, an adapted Pick-Place-Carry scenario is used as test-bed. The robot and the user are in a virtual flat with three rooms, in which there are different kinds of objects varying in terms of colour, type, and position (e.g. blue mug on the kitchen table, red book on the living room table, etc.). The user interacts with the robot using unconstrained speech (Large Vocabulary Speech Recognition) and pointing gestures to ask the robot to perform some specific object manipulation tasks (e.g. move the blue mug from the living room table to the kitchen table). The multimodal dialogue is used to solve ambiguities and to request missing information until task completion (i.e. full command execution) or failure (i.e. explicit user disengagement or wrong command

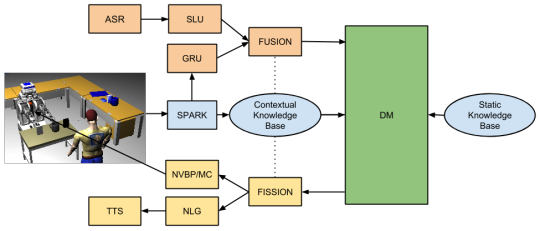


Fig. 3 Architecture of the multimodal and situated dialogue system.

execution). In this study, we specifically focus on tasks where divergent beliefs are prone to be generated as in the Sally and Anne test: a previous interaction has led the user to think that a specific object O is located at A which is out of her view, and an event has changed the object position from A to B without user's awareness. For example, a change performed by another user (or by the robot) without the presence of the first one. Thereby, if the user currently wants to perform a manipulation involving O she may do so using her own believed value (A) of the position property in her communicative act.

Concerning the simulation, the setup of [12] is applied to enable a rich multimodal HRI. Thus, the open-source robotics simulator MORSE [5] is used which provides a realistic rendering through the Blender Game Engine, a wide range support of middleware (e.g. ROS, YARP), and proposes reliable implementations of realistic sensors and actuators which ease the integration on real robotic platforms. It also provides the operator with an immersive control of a virtual human avatar in terms of displacement, gaze, and interactions on the environment, such as object manipulation (e.g. grasp/release an object). This simulator is tightly coupled with the multimodal dialogue system, with the overall architecture given in Figure 3.

In the chosen architecture, the Google Web Speech API[1] for Automatic Speech Recognition (ASR) is combined with a custom-defined grammar parser for Spoken Language Understanding (SLU). The spatial reasoning module, SPARK, is responsible for both detecting the user gestures and generating the per-agent spatial facts (see Sec. 2) used to dynamically feed the contextual knowledge base and allowing the robot to reason over different perspectives of the world. Furthermore, we also make use of a static knowledge base containing the list of all available objects (even those not perceived) and their related static properties (e.g. color). The Gesture Recognition and Understanding (GRU) module catches the gesture-events generated by SPARK during the course of the interaction. Then, a rule-based fusion engine, close to the one presented in [8], temporally aligns the monomodal inputs (speech and gesture) and merges them to convey the list of possible fused inputs to the POMDP-based DM, with speech considered as the primary modality.

The DM implements the extended HIS framework described in Sec. 3. For the reinforcement learning setup, the sample-efficient KTD-SARSA RL algorithm [4] in combination with the Bonus Greedy exploration scheme enables online learning of dialogue strategy from scratch, as in [6]. A reward function is defined to penalise the DM by -1 for each dialogue turn and give it a $+20$ if the right command is performed at the end of the interaction, 0 otherwise. To convey the DM action back

[1] https://www.google.com/intl/en/chrome/demos/speech.html

to the user, a rule-based fission module is employed that splits the high level DM decision into verbal and non-verbal actions. The robot speech outputs are generated by chaining a template-based Natural Language Generation (NLG) module, which converts the sequence of concepts into text, to a Text-To-Speech (TTS) component based on the commercial Acapela TTS system[2]. A Non-verbal Behaviour Planning and Motor Control (NVBP/MC) module produces robot postures and gestures by translating the non-verbal actions into a sequence of abstract actions such as *grasp*, *moveTo*, *release* which are then executed in the simulated environment.

In this study we intend to assess the benefit of introducing the divergent belief management into the multimodal situated dialogue management problem. Thereby, the scenarios of interest require some situations of divergent beliefs between the user and the robot. In real setup those scenarios often need a long term interaction context tracking. To bypass this time consuming process in our evaluation setup, we directly propose a corrupted goal to the user at the beginning of her interaction. So, a false belief about the location value was automatically added concerning an object not visible from the human point of view. Although the situation is artificially generated, the same behaviour can be obtained with the spatial reasoner if the robot performs an action in self-decision mode, or if another human corrupts the scene. Thereby, this setup was used to evaluate the robot's ability to deal with both classical (CLASSIC) and false belief (FB) object manipulation tasks. To do so, we compare the belief-aware learnt system performance (noted BA-LEARNT hereafter) to an handcrafted one (noted BA-HDC), and with two other similar systems with no perspective taking ability (noted LEARNT and HDC respectively). The handcrafted policies make use of expert rules based on the information provided by the summary state to pick the next action to perform (deterministic). They are not considered as the best possible handcrafted policies but as robust enough to manage correctly an interaction with real users. The learnt policies were trained in an online learning settings using a small set of 2 expert users which first performed 40 dialogues without FB tasks and 20 more as a method-specific adaptation (LEARNT with CLASSIC tasks vs BA-LEARNT with FB tasks). In former works we have shown the possibility to learn efficient policies with few tens of dialogue samples, due to expert users better tolerance to poor initial performance combined with more consistent behaviours during interactions [7].

In the evaluation setup, 10 dialogues for the four proposed system configurations (the learnt policies were configured to act greedily according to the value function) were recorded from 6 distinct subjects (2 females and 4 males, around 25yo on average) who interacted with all configurations (within-subjects study), so 240 dialogues in total. 30% of the performed dialogues involve FB tasks. No user had knowledge of the current system configurations and they were proposed in random order to avoid any prior effect. At the end of each interaction, users evaluated the system in terms of task completion with an online questionnaire.

[2] http://www.acapela-group.com/index.html

5 Results

	HDC			BA-HDC			LEARNT			BA-LEARNT		
TASK	*Avg.R*	*Length*	*SuccR*	*Avg.R*	*Length*	*SuccR*	*Avg.R*	*Length*	*SuccR*	*Avg.R*	*Length*	*SuccR*
CLASSIC	14.33	4.81	0.85	14.28	4.86	0.86	17.62	2.95	0.93	17.69	2.88	0.93
FB	9.78	6.67	0.72	13.05	5.61	0.83	12.72	5.94	0.83	13.89	4.78	0.83
ALL	12.97	5.36	0.82	13.92	5.08	0.85	16.15	3.85	0.9	16.55	3.45	0.9

Table 1: System performance on classic (CLASSIC), false belief (FB) and all (ALL) tasks in terms of average cumulative discounted reward (Avg.R), average dialogue length in terms of system turns (Length) and average success rate (SuccR).

Table 1 is populated with the performance obtained by the four system configurations discussed above considering CLASSIC and FB tasks. These results are first given in terms of mean discounted cumulative rewards (Avg.R). According to the reward function definition, this metric expresses in a single real value the two variables of improvement, namely the success rate (accuracy) and the number of turns until dialogue end (time efficiency). However, both metrics are also presented for convenience. The results in Table 1 were gathered in test condition where no exploration of the RL method is allowed. Thus, they basically consist of a mere average over the 60 performed dialogues for each method and metric.

The differences observed between the LEARNT/BA-LEARNT and the HDC/BA-HDC on the overall performance (row ALL) shows the interest of considering RL methods rather than handcrafted policies. Indeed, only 60 training dialogues are enough to outperform both handcrafted solutions. On CLASSIC tasks the performance between LEARNT and BA-LEARNT as well as between HDC and BA-HDC must be considered similar. Thus, the divergent belief resolution mechanism doesn't seem to impact the dialogue management when divergent belief situations do not appear. For BA-HDC this statement could be expected (in lack of false belief, the rules are the same as HDC). However for BA-LEARNT the tested policy is learnt and the action assignment process is optimized with an additional degree of complexity (larger state/action space than in LEARNT), so a loss could have been observed. The performances between LEARNT and BA-LEARNT and respectively between HDC and BA-HDC on FB tasks appear in favour of the BA-systems (both show a higher success rate and a slightly more time efficient dialogue management process - average gain of 1 turn). However the quantitative comparison between the system configurations is not ensured to be relevant due to the relatively high confidence interval on considered metrics (e.g. success rate confidence interval for row FB is around 0.2 for all system configurations). Two main reasons account for this status quo. First, a limited amount of observations involving the different system configurations (due to experimental cost). Second, the expected marginal gain in terms of the considered metrics. Indeed, the current system is learnt on some overall task completion and efficiency criterion. However solving divergent belief situations in a pick and place scenario can not be considered a critical factor influencing these criterion greatly but just a way to cope with an additional (not dominant) degree of

uncertainty and to improve user experience and naturalness of the interaction with the embodied agent.

R_1: Can I help you? U_1: Bring me the book on my bedside table

R_2:	**The brown book is not on the bedside table, it has been moved to the kitchen.**
U_2:	Ok, bring it to me
R_3:	I will give you the brown book that is on the kitchen table

(a)

R_2:	Do you want the red one?
U_2:	No, the brown book.
R_3:	There is no brown book in your bedroom but there is one in the kitchen
U_3:	Are you sure? Well, bring me that one.
R_4:	I will give you the brown book that is on the kitchen table

(b)

Table 2: Dialogue examples with (a) and without (b) divergent belief reasoning in the case of an unknown (from the user's point of view) interchange between a red and a brown book.

To have better insights on what the main differences between the four dialogue strategies are we also performed a qualitative study. In this study we precisely identify the behavioural differences due to introducing a FB handling mechanism in a learning setup. Overall, it is observed that confirmation acts (e.g. confirm, offer) are more accurate and less frequent for the two learnt methods. For instance, when the learnt systems are confident on the top object manipulation hypothesis they predominantly performed the command directly rather than trying to check its validity further as in the handcrafted versions. In Table 2 two dialogue samples extracted from the evaluation dataset illustrate the differences between non-BA and BA dialogue management on the same FB task (here a red book was interchanged with a brown one). If the belief divergence problem is not explicitly taken into account (as in (a)) the DM can be constrained to deal with an additional level of misunderstanding (see (b) from R_2 to U_3). We can also see in (b) that the non-BA system was able to succeed FB tasks (explaining the relative high LEARNT performance on FB tasks). Indeed, if the object is clearly identified by the user (e.g. color and type) the system can release the constraint of the false position and thus is able to make an offer on (execute) the "corrected" form of the command involving the true object position. Concerning the main differences between BA-LEARNT and BA-HDC, we observed a less systematic usage of the *inform divergent belief* act in the learnt case. BA-LEARNT first tries to reach a high confidence on the true presence of the object involved in the belief divergence in the user goal. Furthermore, BA-LEARNT, like LEARNT, has learnt alternative mechanisms to fulfil FB tasks such as direct execution of the user command (which also avoids misunderstanding) when the convoyed piece of information seems to be sufficient to identify the object.

6 Conclusion

In this paper, we described how a user belief realtime tracking framework can be used along with a multimodal POMDP-based dialogue management. The evaluation of the proposed method with real users confirms that this additional information helps to achieve more efficient and natural task planning (and does not harm

handling of normal situations). Our next step will be to integrate the multimodal dialogue system on the robot and carry out evaluations in real setting to uphold our claims in an fully realistic configuration.

Acknowledgements This work has been partly supported by the French National Research Agency (ANR) under project reference ANR-12-CORD-0021 MaRDi.

References

1. F. U. Baron-Cohen S, Leslie AM. Does the autistic child have a 'theory of mind'? *Cognition*, 21(1):37 – 46, 1985.
2. C. Breazeal, M. Berlin, A. Brooks, J. Gray, and A. Thomaz. Using perspective taking to learn from ambiguous demonstrations. *Robotics and Autonomous Systems*, 2006.
3. C. Breazeal, J. Gray, and M. Berlin. An embodied cognition approach to mindreading skills for socially intelligent robots. *I. J. Robotic Res.*, 2009.
4. L. Daubigney, M. Geist, S. Chandramohan, and O. Pietquin. A comprehensive reinforcement learning framework for dialogue management optimization. *Journal on Selected Topics in Signal Processing*, 6(8):891–902, 2012.
5. G. Echeverria, N. Lassabe, A. Degroote, and S. Lemaignan. Modular open robots simulation engine: Morse. In *ICRA*, 2011.
6. E. Ferreira and F. Lefevre. Expert-based reward shaping and exploration scheme for boosting policy learning of dialogue management. In *ASRU*, 2013.
7. E. Ferreira and F. Lefèvre. Social signal and user adaptation in reinforcement learning-based dialogue management. In *Proceedings of the 2nd MLIS Workshop*, pages 61–69. ACM, 2013.
8. H. Holzapfel, K. Nickel, and R. Stiefelhagen. Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3d pointing gestures. In *ICMI*, 2004.
9. L. Kaelbling, M. Littman, and A. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence Journal*, 101(1-2):99–134, May 1998.
10. E. Levin, R. Pieraccini, and W. Eckert. Learning dialogue strategies within the markov decision process framework. In *ASRU*, 1997.
11. L. Lucignano, F. Cutugno, S. Rossi, and A. Finzi. A dialogue system for multimodal human-robot interaction. In *ICMI*, 2013.
12. G. Milliez, E. Ferreira, M. Fiore, R. Alami, and F. Lefèvre. Simulating human robot interaction for dialogue learning. In *SIMPAR*, pages 62–73, 2014.
13. G. Milliez, M. Warnier, A. Clodic, and R. Alami. A framework for endowing interactive robot with reasoning capabilities about perspective-taking and belief management. In *ISRHIC*, 2014.
14. F. Pinault and F. Lefèvre. Unsupervised clustering of probability distributions of semantic graphs for pomdp based spoken dialogue systems with summary space. In *KRPDS*, 2011.
15. N. Roy, J. Pineau, and S. Thrun. Spoken dialogue management using probabilistic reasoning. In *ACL*, 2000.
16. R. Sutton and A. Barto. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, 9(5):1054–1054, 1998.
17. B. Thomson and S. Young. Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems. *Computer Speech and Language*, 24(4):562–588, 2010.
18. J. Trafton, N. Cassimatis, M. Bugajska, D. Brock, F. Mintz, and A. Schultz. Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man, and Cybernetics*, 35(4):460–470, 2005.
19. B. Tversky, P. Lee, and S. Mainwaring. Why do speakers mix perspectives? *Spatial Cognition and Computation*, 1(4):399–412, 1999.
20. H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103 – 128, 1983.
21. 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, 2010.