Towards Action Selection Under Uncertainty for a Socially Aware Robot Bartender

Mary Ellen Foster
M.E.Foster@hw.ac.uk

Simon Keizer
S.Keizer@hw.ac.uk

Oliver Lemon
O.Lemon@hw.ac.uk

The Interaction Lab, School of Mathematical and Computer Sciences
Heriot-Watt University, EH14 4AS, Edinburgh, UK

ABSTRACT

We describe how the state representation of a socially aware robot is being extended to handle uncertainty. It incorporates the full range of information provided by the input sensors, including the confidence of all hypotheses. We also show how the Interaction Manager is being updated to make use of the extended representation.

Categories and Subject Descriptors: I.2.9 [Artificial intelligence]: Robotics – Operator interfaces

Keywords: Belief tracking; Social robotics

1. INTRODUCTION

A crucial aspect of the design of an interactive multimodal system is state management: transforming the noisy, continuous hypotheses produced by the low-level input processing components into a form that can be used as the basis for higher-level action selection. A state representation that considers only the highest-confidence input hypotheses is straightforward to maintain and reason with, but discards a great deal of potentially useful information. On the other hand, a representation that takes into account the full set of input hypotheses—along with their estimated confidence scores—can be more robust and informative, but requires more sophisticated methods of maintenance and more complex forms of reasoning. This general problem of state estimation under uncertainty is a multimodal version of the *belief tracking* problem for spoken dialogue systems [8, 9]: sequences of uncertain observations are processed in order to build up an increasingly accurate *belief state* regarding user goals as the interaction progresses.

We are currently addressing this issue in the context of the JAMES robot bartender (Figure 1), which has the goal of supporting socially appropriate multi-party interaction in a bartending scenario. Based on (uncertain) observations about the users in the scene provided by the vision and speech recognition components, the system maintains a model of the social context, and decides on effective and socially appropriate responses in that context. In this paper, we describe how the state representation initially developed for the JAMES robot is being extended to incorporate the full range of information from the input sensors, and how the Interaction Manager is being updated to make use of that extended representation.

Figure 1: The JAMES bartender

2. REPRESENTING UNCERTAINTY

Within the JAMES system, the task of the Social State Recogniser (SSR) is to turn the continuous stream of messages produced by the low-level input and output components into a discrete representation of the world, the robot, and all entities in the scene, integrating social, interaction-based, and task-based properties. In addition to storing all of the low-level sensor information, the SSR also infers additional relations that are not directly reported by the sensors. For example, it fuses information from vision and speech to determine which user should be assigned to a recognised speech hypothesis, and uses the vision information to estimate whether each customer is currently seeking attention from the bartender [1].

For the initial version of the SSR, the state was represented as a single list of properties and their values, and a fixed confidence threshold was used to decide whether to include any given property into the state. Details of this representation are given in [5], and a sample state excerpt is shown in the highlighted portion of Table 1. This initial representation has since been extended to incorporate two features that the previous version did not exhibit:

- Every relation in the state has an associated **confidence** value, represented as a number between 0 and 1.
- Every relation in the state can potentially have **multiple values**, with each possible value having its own confidence value.

Table 1 shows a full state using this representation. In addition to the old-style state information in the highlighted portion, this state also includes confidence scores on all relations—meaning that the low-confidence lastAct(A1) relation can now be included—and also shows multiple possible values for the drink_order(A1) relation.

Incorporating multiple hypotheses and confidence scores into the state requires additional processing in the SSR. For the vision data, we make use of the information from the JAMES computer vision system [4], which provides a continuous estimate of the location,

seeksAttention(A1)	true	0.75
seeksAttention(A2)	false	0.45
lastSpeaker()	A1	1.0
lastEvent()	userSpeech(A1)	1.0
drink_order(A1)	green lemonade	0.677
	blue lemonade	0.322
lastAct(A1)	greet	0.25

Table 1: State excerpt, showing both the previous representation (highlighted portion) and the new representation

gaze behaviour, and body language of all people in the scene in real time. Every feature reported by the vision system includes an estimated confidence value; these values are incorporated into the state, and also used to determine the confidence value for derived properties such as seeksAttention.

For speech recognition, we make use of the Microsoft Kinect for Windows API [3], which produces an *n*-best list of recognition hypotheses, each with an estimated confidence score, along with an estimate of the sound source angle and the angle confidence. The recognised hypotheses are parsed to extract the syntactic and semantic information using a grammar implemented in OpenCCG [7], while the source angle is used together with the location information from vision to estimate which of the customers in the scene is most likely to have been speaking. If a possible speaker is found, the semantic information from speech is used to update the lastAct relation. In the case that the customer says something regarding their drink order, we also update the value of the drink_order relation, using the generic belief tracking procedure proposed by Wang and Lemon [6], which maintains beliefs over user goals based on a small number of domain-independent rules, using basic probability operations. This allows us to maintain a dynamically-updated list of the possible drink orders made by each customer in the scene, with an associated confidence value for each order.

3. REASONING WITH UNCERTAINTY

Within JAMES, the Social Skills Executor (SSE) controls the behaviour of the robot system, based on the state updates it receives from the SSR. In previous work, we have shown that strategies for action selection in this multi-party interactive context can be learnt automatically in interaction with a Multi-User Simulated Environment (MUSE) [2]. Using a hierarchical MDP (Markov Decision Process) model, an action selection policy was trained, mapping combinations of state features to the actions that give the highest expected cumulative reward. So far, these policies have been based on features extracted from the previous SSR state representation, which—as described above—does not take into account any information about uncertainty or confidence. In ongoing work, we are extending this approach to produce a model which includes features representing uncertainty information, such as the probability of the top hypothesis, the entropy of the state distribution, and the *n*-best list of possible drink orders and their confidence scores. The set of possible system actions is being extended to include dialogue acts for clarifying drink orders (e.g., "Did you say 'blue lemonade'?"), as well as clarifying whether or not a customer wants to order in the first place (e.g., "Did you want to order?"). During training, the system should learn automatically when to take the time to ask users for clarifications in order to increase its confidence in the customer's goals, and when to trust its current belief about the customer's goals and proceed with the task—risking lower user satisfaction if that belief turns out to be wrong and needs to be repaired.

In order to learn SSE policies for the system to behave in a manner that is not only socially appropriate, but also robust to the uncertainty arising from noisy audio-visual input, the simulated environment is also being extended. In addition to simulating the behaviour of multiple customers, we are also incorporating an error model based on previously recorded data that produces realistically noisy input signals for the SSR to process. The state updates (including uncertainty information) sent to the SSE based on this simulated input will then correspond more accurately with the updates the SSE can expect when interacting with real users, so the learnt policy is thus more likely to result in better performance in practice.

4. SUMMARY AND FUTURE WORK

We have described how the state representation used in a robot bartender is being extended to deal with various forms of uncertainty about the sensed state, and have shown how the Interaction Manager is being extended to make use of the updated state information. Our approach is similar to that taken in previous spoken dialogue systems, where it has been shown that a combination of belief tracking and reinforcement learning, most notably in the form of POMDP (Partially Observable Markov Decision Process) approaches, leads to more robust system behaviour [10]. In our work, however, we aim to deal with noise not just in the speech hypotheses, but also in the multimodal information provided by the vision system.

We are currently carrying out a user study comparing a version of the bartender that deals with all of the above forms of uncertainty to one that does not. Based on the behaviour of previous versions of the bartender—which did not incorporate any of the uncertainty in the state—we expect to see a positive impact in task performance (i.e., the number of drinks correctly served), since the bartender should clarify lower-confidence or ambiguous state hypotheses before moving on to serving the drinks.

5. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 270435, JAMES: Joint Action for Multimodal Embodied Social Systems (`james-project.eu`).

6. REFERENCES

[1] M. E. Foster, A. Gaschler, and M. Giuliani. How can I help you? Comparing engagement classification strategies for a robot bartender. In *Proceedings of ICMI*, 2013. doi:10.1145/2522848.2522879.

[2] S. Keizer, M. E. Foster, O. Lemon, A. Gaschler, and M. Giuliani. Training and evaluation of an MDP model for social multi-user human-robot interaction. In *Proceedings of SIGdial*, 2013.

[3] Microsoft Corporation. Kinect for Windows. URL `http://www.microsoft.com/en-us/kinectforwindows/`.

[4] M. Pateraki, M. Sigalas, G. Chliveros, and P. Trahanias. Visual human-robot communication in social settings. In *Proceedings of ICRA Workshop on Semantics, Identification and Control of Robot-Human-Environment Interaction*, 2013.

[5] R. P. A. Petrick and M. E. Foster. Planning for social interaction in a robot bartender domain. In *Proceedings of ICAPS*, 2013.

[6] Z. Wang and O. Lemon. A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. In *Proceedings of SIGDial*, 2013.

[7] M. White. Efficient realization of coordinate structures in Combinatory Categorial Grammar. *Research on Language and Computation*, 4 (1):39–75, 2006. doi:10.1007/s11168-006-9010-2.

[8] J. Williams, A. Raux, D. Ramachandran, and A. Black. The dialog state tracking challenge. In *Proceedings of SIGDial*, 2013.

[9] J. D. Williams. A belief tracking challenge task for spoken dialog systems. In *NAACL HLT 2012 Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data*, 2012.

[10] S. Young, M. Gašić, S. Keizer, F. Mairesse, 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. doi:10.1016/j.csl.2009.04.001.