On-Line Learning of Lexical Items and Grammatical Constructions via Speech, Gaze and Action-Based Human-Robot Interaction

Gregoire Pointeau, Maxime Petit, Xavier Hinaut, Guillaume Gibert, Peter Ford Dominey

On-Line Learning of Lexical Items and Grammatical Constructions via Speech, Gaze and Action-Based Human-Robot Interaction

Grégoire Pointeau, Maxime Petit, Xavier Hinaut, Guillaume Gibert, Peter Ford Dominey

Robot Cognition Laboratory
INSERM Stem Cell and Brain Research Institute, Lyon, France
firstname.name@inserm.fr

Abstract

In order to be able to understand a conversation in interaction, a robot, has to first understand the language used by his interlocutor. A central aspect of language learning is adaptability. Individuals can learn new words and new grammatical structures. We have developed learning methods that allow the humanoid robot iCub to robot can learn new lexical items by interaction with the human and consolidation of its autobiographical memory. Then, based on these open class words, the robot can bootstrap the acquisition of novel grammatical structures in real-time. Finally, we demonstrate how human gaze can be monitored, and could be used in order to reduce referential ambiguity inherent in such learning conditions. These learning capabilities are demonstrated in a collection of videos.

1. Learning of "*open-class*" word through Human Robot interaction, using an autobiographical-like memory

In robotic systems based on human-inspired robot task learning [1], significant attention has been allocated to the mechanisms that underlie the ability to acquire knowledge from and encode the individual's accumulated experience [2], and to use this accumulated experience to adapt to novel situations [3]. In this context we consider research on autobiographical memory (ABM) and mechanisms by which ABM can be used to generate new knowledge [4].

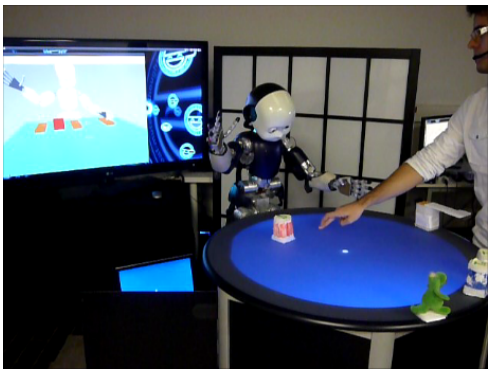

Figure 1: Physical interaction between the iCub and an human. Both are interacting with the ReacTable (the blue table). Behind the robot, its internal representation is displayed on a screen.

The objective of the current research is to demonstrate how an autobiographical memory system, coupled with mechanisms for detecting and extracting regularities can be used to construct a progressive hierarchy of spatial, and temporal relations that provide the basis for learning and executing shared plans.

The mixed-initiative dialogue between the Robot and the Agent is done as described in the Figure 1 and 2, by a simple oral interaction.

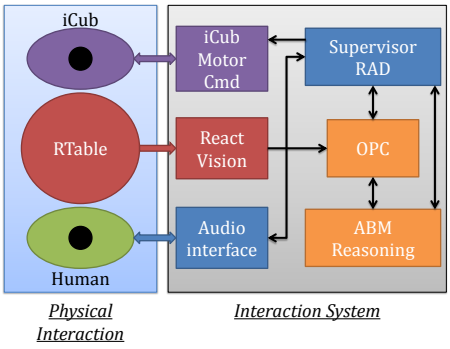

Figure 2: System Architecture Overview. Human and iCub interact face-to-face across the ReacTable, which communicates object locations via ReactVision to the Object Property Collector (OPC). The Supervisor coordinates spoken language and physical interaction with the iCub via spoken language technology in the Audio interface. The autobiographical memory system encodes world states and their transitions due to human and robot action as encoded in the OPC.

The main purpose of this autobiographical-like is to store and to manage information related to the world, with the help of a dialogue between the Human and the Robot. The dialogue will be about the object of focus, the action to perform and what to be attentive to.

2. Learning the mapping between structure of sentences and the structure of meaning - what we called "*grammatical constructions*"

The goal of this language model is to provide a real-time and adaptive spoken langue interface between humans and a humanoid robot. The system should be able to learn new grammatical constructions in real-time, and then use them immediately following or in a later interactive session. In order

to achieve this we use a recurrent neural network (RNN) of a few hundred neurons (from 100 to 500) with the paradigm of Echo State Network [5].

The model processes sentences as grammatical constructions [6] – e.g. *"put the toy on the left"* – and meanings in a predicate form predicate (agent, object/location) – e.g. put (toy, left). A grammatical contruction is an abstraction of a sentence, in which the open class words (e.g. nouns and verbs : put, grasp, toy, left, right ...) are replaced by a common SW (Semantic Word) marker. Thus the RNN only have information about the remaining words in the sentence, namely close class words (prepositions, auxiliary verbs, etc. : the, on, to, is ...). Therefore the model is able to deal with sentences that have the same constructions than previously seen sentences.

The target behavior of the system is to learn two conditions (see Figure 3). In action performing (AP), the system should learn to generate the proper robot command, given a spoken input sentence. In scene description, the system should learn to describe scenes given the extracted spatial relation (SR). More details on the neural model processing could be find in [7] and [8].

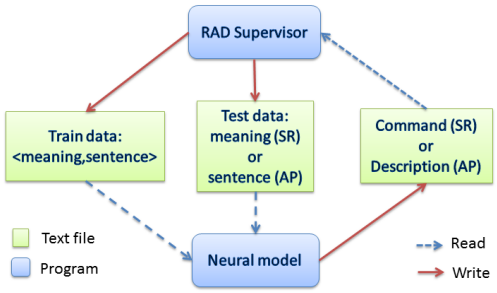

Figure 3: Communication between the speech recognition tool (Supervisor RAD) – that also controls the robotic platform – and the neural model.

Training corpus for the neural model can be generated by the interaction with the user teaching the robot by describing spatial relations or actions, creating <sentence, meaning> pairs.

In the AP condition, we demonstrate that the model can learn and generalize to complex sentences including *"Before you put the toy on the left point the drums."*; the robot will first point the drums and then put the toy on the left: showing here that the network is able to establish the proper chronological order of actions.

Likewise, in the SR condition, the system can be exposed to a new scene and produce a description such as "To the left of the drums and to the right of the toy is the trumpet."

3. The importance of gaze and shared-attention

The eyes and more specifically the gaze is an important signal for social communication. Even at the early stage of the interaction with an avatar, the initiation of contact, it plays a crucial role [9]. There are also evidences that at the neural level, direct gaze elicit greater electrophysiological responses (N170 and early posterior negativity potentials) than averted gaze [10] in live condition. Several factors influence eye gaze pattern during a dyadic interaction. Environmental conditions such as noise in the auditory channel [11] and the role in the interaction or the cognitive state [12] modify the gaze pattern. A set of principal communicative speech and gaze cues in human-human interaction were identified and then formalized and implemented in a humanoid robot [13]. The authors demonstrated the pertinence of these cues and showed that gaze significantly facilitates cooperation as measured by human reaction time in a cooperative game. What are less clear are the limits of acceptability of spatial and temporal disruption of the interlocutor's gaze. To answer these questions, we developed a super Wizard of Oz platform. In this platform, a confederate's head movements and gaze trajectories are tracked in real-time and replicated on the iCub humanoid robot. By manipulating the latency and spatial shifting of head movements and/or gaze trajectories, we will determine the real limits that are acceptable for humans during face to face interaction.

4. Conclusion

The ability of the system to control gaze and monitor the gaze of the human will be of central importance in establishing a shared space for learning and interaction. The ensemble of mechanisms described for learning open class words, using those words to bootstrap grammar acquisition, and the use of gaze for shared attention will be crucial in the further development of speech-based interaction with robots. Demonstration videos of these functions can be seen at the following addresses:

- Gaze impact:
  http://www.youtube.com/watch?v=Sv2hQCW7q54
- Autobiographical memory, learning through interaction:
  http://www.youtube.com/user/MaximePetitU846/videos
- Understanding complex sentence structure:
  http://youtu.be/AUbJAupkU4M
- Producing complex sentence structure:
  http://youtu.be/3ZePCuvygi0

5. References

[1] Wu, X. and J. Kofman, *"Human-inspired robot task learning from human teaching. in Robotics and Automation"*, 2008. ICRA 2008. IEEE International Conference on. 2008. IEEE.

[2] Broz, F., et al., *"Learning behavior for a social interaction game with a childlike humanoid robot."*, in Social Learning in Interactive Scenarios Workshop, Humanoids. 2009.

[3] Pointeau, G., M. Petit, and P.F. Dominey, *"Robot Learning Rules of Games by Extraction of Intrinsic Properties."*, in ACHI 2013, The Sixth International Conference on Advances in Computer-Human Interactions. 2013.

[4] Conway, M.A. and C.W. Pleydell-Pearce, *"The construction of autobiographical memories in the self-memory system."*, Psychol Rev, 2000. 107(2): p. 261-88.

[5] H. Jaeger, *"The echo state approach to analysing and training recurrent neural networks"*, Tech. Rep. GMD Report 148, German National Research Center for Information Technology, 2001.

[6] Goldberg, A., *"Constructions: A Construction Grammar Approach to Argument Structure. Cognitive Theory of Language and Culture"*, ed. G. Fauconnier, G. Lakoff, and E. Sweetser, 1995, Chicago: University of Chicago Press. 265.

[7] Hinaut, X (2013). *"Recurrent Neural Networks for Abstract Sequence and Grammatical Structure Processing, with an Application to Human-Robot Interaction."*, PhD thesis. Universit Claude Bernard Lyon 1.

[8] Hinaut X, Dominey PF (2013) *"Real-Time Parallel Processing of Grammatical Structure in the Fronto-Striatal System: A Recurrent Network Simulation Study Using Reservoir Computing."*, PLoS ONE 8(2): e52946. doi:10.1371/journal.pone.0052946

[9] N. Bee, E. Andre, and S. Tober, *Breaking the Ice in Human-Agent Communication: Eye-Gaze Based Initiation of Contact with an Embodied Conversational Agent"*, presented at the Proceedings of the 9th International Conference on Intelligent Virtual Agents, Amsterdam, The Netherlands, 2009.

[10] L. M. Ponkanen, A. Alhoniemi, J. M. Leppanen, and J. K. Hietanen, *"Does it make a difference if I have an eye contact with you or with your picture? An ERP study"*, Social Cognitive and Affective Neuroscience, 2010.

[11] E. Vatikiotis-Bateson, I. M. Eigsti, S. Yano, and K. G. Munhall, *"Eye movement of perceivers during audiovisual speech perception"*, Perception & Psychophysics, vol. 60, pp. 926-940, Aug 1998.

[12] G. Bailly, S. Raidt, and F. Elisei, (2010, Gaze, conversational agents and face-to-face communication. Speech Communication [doi:DOI:10.1016/j.specom.2010.02.015]. 52(6), 598-612.

[13] J. Boucher, J. Ventre-Dominey, P. F. Dominey, S. Fagel, and G. Bailly, *"Facilitative Effects of communicative gaze and speech in Human-Robot Cooperation"*, presented at the International Workshop on Affective Interaction in Natural Environments, 2010.