Online Language Learning to Perform and Describe Actions for Human-Robot Interaction

Xavier Hinaut[1,2], *Maxime Petit*[1,2], *Peter F. Dominey*[1,2]

[1] Stem cell and Brain Research Institute, INSERM U846, 18 Avenue Doyen Lepine, 69500 Bron, France

[2] Université de Lyon, Université Lyon I, 69003, Lyon, France

{xavier.hinaut ; maxime.petit ; peter.dominey}@inserm.fr

Abstract

The goal of this research 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 of 500 neurons - echo state network with leaky neurons [1].

The model processes sentences as grammatical constructions, in which the semantic words (nouns and verbs) are extracted and stored in working memory, and the grammatical words (prepositions, auxiliary verbs, etc.) are inputs to the network. The trained network outputs code the role (predicate, agent, object/location) that each semantic word takes. In the final output, the stored semantic words are then mapped onto their respective roles. The model thus learns the mappings between the grammatical structure of sentences and their meanings.

The humanoid robot is an iCub [2] who interacts around a instrumented tactile table (ReacTable™) on which objects can be manipulated by both human and robot. A sensory system has been developed to extract spatial relations. A speech recognition and text to speech off-the-shelf tool allows spoken communication. In parallel, the robot has a small set of actions (put(object, location), grasp(object), point(object)). These spatial relations, and action definitions form the meanings that are to be linked to sentences in the learned grammatical constructions.

The target behavior of the system is to learn two conditions. In action performing (AP), the system should learn to generate the proper robot command, given a spoken input sentence. In scene description (SD), the system should learn to describe scenes given the extracted spatial relation.

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. It could also be edited by hand to avoid speech recognition errors. These interactions between the different components of the system are shown in the Figure 1.

The neural model processes grammatical constructions where semantic words (*e.g. put*, *grasp*, *toy*, *left*, *right*) are replaced by a common marker. This is done with only a predefined set of grammatical words (*after*, *and*, *before*, *it*, *on*, *the*, *then*, *to*, *you*). Therefore the model is able to deal with sentences that have the same constructions than previously seen sentences.

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 SD 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."

In future research we can exploit this learning system in the context of human language development. In addition, the neural model could enable errors recovery from speech to text recognition.

Index Terms: human-robot interaction, echo state network, online learning, iCub, language learning.

References

[1] 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.

[2] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The iCub humanoid robot: an open platform for research in embodied cognition," in *PerMIS: Performance Metrics for Intelligent Systems Workshop*, Washington DC, USA, pp. 19-21, 2008.

Acknowledgements

This work has been financed by the FP7-ICT 231267 Project Organic and by the FP7-ICT-270490 Project EFAA. Neural model has been developed with Oger toolbox: http://reservoir-computing.org/organic/engine.

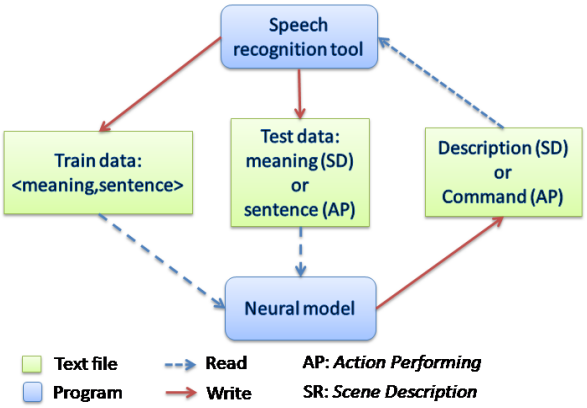

Figure 1: Communication between the speech recognition tool (that also controls the robotic platform) and the neural model.