Towards language acquisition in autonomous robots

Vadim Tikhanoff[1], Angelo Cangelosi[1], Jun Tani[2], Giorgio Metta[3]

[1] University of Plymouth, UK
[2] RIKEN Brain Science Institute, Japan
[3] Italian Institute of Technology, Italy

Abstract

In this paper we present a novel cognitive robotic model to study language acquisition in autonomous robots through the grounding of words in sensorimotor representations. The aim of this new model is to extend previous work on language grounding in simulated cognitive agents (Cangelosi & Riga 2006) to the new robotic platform iCub (Metta et al. 2006). The language learning model is based on the use of artificial neural networks controllers (Kohonen 1995; Tani 2004).
The model proposed here is based on a series of interconnected modules to gather and integrate visual and linguistic information for a language comprehension task. The model comprises of a vision module, a sound perception and feature extraction module and a language integration and recognition network.
The vision acquisition module takes input from the robot's cameras and applies approximation techniques for the purpose of detecting shapes, size and colour features of individual objects. The classification of a spoken word is based on the sequence of the most activated neurons of a self organizing map (SOM) with a 10 x 10 topological 2D grid. The SOM model has been trained on 112 English words and 544 syllable utterances both from two different speakers, for determining the ability of the system to distinguish between all words.
The language integration module is based on a Recurrent Neural Network with Parametric Biases (RNNPB). This network is particularly suitable for online learning of behaviour in robots. Two experiments were carried out to test the language learning model. The first consists in the recognition and classification of the speech signals as an imitation task without the integration of the vision module. This experiment has been based on the use of 20 words. Each training patterns (words) consists of a sequence of x/y coordinates of the SOM map. During the interaction phase of RNNPB training, the system learns to imitate the SOM word feature outputs pattern by predicting their next pattern. The network successfully learns to recognize spoken words with a final mean square error of the output nodes of 0.082.
The second experiment consists of the integration of the vision and speech modules for learning and grounding of the names of objects. This experiment uses as input stimuli the combination of the features extracted from the visual module and the SOM output patterns. The output units predict the SOM sequence for the object name shown in the picture. The final square error of the output nodes was 0.003 over all the learning results. The model was able to categorize and name two objects which share some features (e.g. shape) but differ in other dimensions (e.g. colour).

This preliminary work demonstrates the successful integration of a SOM network to classify spoken words with the RNNPB network capable of on-line learning and naming of visual objects.
Future plans to extend the model include the learning of motor responses to be associated to the visual input of different objects and the capability to combine groups of words to describe visual scenes involving multiple objects.