Grounding Words in Actions: Experiments with the iCub Humanoid Robot

Davide Marocco and Angelo Cangelosi
Centre for Robotics and Neural Systems, School of Computing and Mathematics, University of Plymouth, UK.

Human language is a formidable communication system. It allows to describe the world around us and exchange our thoughts. Among humans language is used in many different ways, such as describing what we perceive, asking others to perform certain actions or simply engage in conversation (Siskind, 2001). At the core of this description there is our ability to understand and use correctly the meaning that words represent. Especially the first two cases above require to ground language in perception and action processes. Moreover, by focusing the attention on our abilities to easily describe dynamics that happen in time and entails specific relations between objects and object properties, the process of grounding language in perception and actions means that, when we describe a given scene or we ask someone to perform a certain action, the words used must be linked with physical entities in the scene or in actions that can be either observed or desired.

In order to understand the link by which words are connected with objects and actions, an interesting recent approach is based on the use of computational and robotics models that are able to show a certain degree of language abilities and several computational models have been proposed to study communication and language in artificial cognitive systems, such as robots and simulated agents (Cangelosi & Parisi 2002; Lyon, Nehaniv & Cangelosi, 2007).

Current literature tends to define nouns as words associated to physical (or even abstract) entities, and verbs as words that represent actions (or, in general, events that happen in time). Grounded computational models so far are mainly focused on grounding nouns on sensorimotor object representations and verbs on actions that are directly performed by the agent. Actions, however, are not only an exclusive domain of the agent. Physical objects in the environment, for example, can also perform actions. Only few studies focus the attention on the acquisition of actions words that are connected to property of objects, such as rolling for a ball, or words that underpin a dynamical and force-varied interaction with objects, such as hit or move.

The aim of the present research is to study how a humanoid robot can learn to understand the meaning of action words (i.e. words that represent dynamical events that happen in time) by physically acting on the environment and linking the effects of its own actions with the behaviour observed on the objects before and after the action. This will allow the agent to give an interpretation of a given scene that develops in time, and is grounded on its own bodily actions and sensory-motor coordination. Object manipulation, therefore, is the central concept behind this research.

To approach the research issue related to the grounding of action words in sensorimotor coordination, we present a robotic model. For the experiments we used a simulation of the iCub humanoid robot (Tikhanoff et al 2008) controlled by an artificial neural network. The robot can interact with three objects located on a desk in front of it (a cube, a ball, and a fix cylinder stuck vertically on the desk) and its neural control system is trained through a “Back Propagation Through Time” algorithm. By manipulating the environment, the robot can learn the association between objects and physical property of such objects. The neural system is a fully recurrent neural network with ten hidden units, height inputs, and height outputs. Activations of input units are divided into four sensory units and three linguistic units. Three of the four sensory units provide information about current angles of three corresponding joints of the right arm. The fourth sensory unit encodes the value of a binary tactile sensor. The three linguistic input units represent a local binary encoding of the three objects. The neural network is trained to learn the sensorimotor contingencies produced by the manipulation of the robot toward the object in the environment. In particular, given an input pattern at time t the network must be able to predict the input pattern at $t+1$. After the training of 10 different neural networks, all the controllers were able to correctly predict the next sensory input state, on the basis of the current input state. Tests shown that an error E smaller that 0.001 produces neural controllers capable of performing the task with an good degree of generalisation. After the run of many different tests, results show that the linguistic input is tightly connected with the sensory-motor dynamic produced by the interaction with the object. Experiments demonstrate the ability of the robot to correctly categorise objects, also in the absence of direct linguist input, and to produce the corresponding linguistic label only on the bases of its sensory-motor state. A further generalisation test conducted with different objects (that have same physical properties of the objects used during the training) shows that the robot is able to correctly categorise not the object itself, but the actions performed by the object, observed during the manipulation (eg. rolling, sliding, being fixed). We believe that the principal contribution of this model is showing that the grounding of action words relies directly to the way in which an agent interacts with the environment and manipulates it. The dynamical properties of external objects, such as being movable, or being fix, are embodied and directly represented in the way in which the agent experiences the reactions produced on its perceptual system by its own, self-generated, active manipulation of the world.