MULTIMODAL AND CROSS-MODAL DISTRIBUTIONAL SEMANTICS: TOWARDS COMMON SEMANTIC SPACE FOR WORDS AND THINGS

Baroni M. (marco.baroni@unitn.it)

Center for Mind/Brain Sciences, University of Trento, Italy

Distributional semantic models (DSMs) capture various aspects of word meaning with vectors summarizing their patterns of co-occurrence in large text corpora, under the assumption that the contexts in which words occur are good cues of what they mean. DSMs have been very successful empirically, and they have been used to model increasingly sophisticated linguistic and cognitive phenomena.

However, current DSMs account for linguistic meaning entirely in terms of linguistic signs (the "meaning" of a word is a summary of the linguistic contexts in which the word occurs). This leads to two big conceptual problems: lack of grounding and lack of reference. Concerning the former, cognitive scientists have accumulated plenty of evidence that, for human beings, meaning is strongly embodied in the sensory-motor system, so a semantic theory that completely dissociates meaning from perception and action is, a priori, a rather implausible model of how humans work — a fact that has also empirical consequences in the surprisingly bad performance of DSMs on simple tasks requiring perceptual information. Lack of reference is perhaps an even more serious problem. A theory that has no way to connect the semantic representation of a linguistic expression to states of the world is clearly missing something fundamental about language, as it has no way to explain how we can talk about things!

Interestingly, in the last decade, it has become common in computer vision to represent images through vectors recording the distribution of automatically extracted discrete visual features in them — a representation that is very similar to the one that DSMs assume for words. This suggests that we might be able to free DSMs from their textual cage by establishing a connection with the visual world by means of such vector-based image-representation techniques.

In my talk, after a brief general introduction to distributional semantics, I will discuss experiments we carried out in the last few years in which we tackle the grounding problem (DSMs with richer multimodal semantic representations that combine linguistic and visual features), and recent work in which we started dealing with the reference issue (how to map images and linguistic expressions across modalities to a common space, in order to link language to the world out there). The case studies I will present include simulating human semantic similarity judgments, predicting the color of objects, modeling brain data and learning names and verbally-expressed attributes of objects present in pictures from indirect evidence.