Mind the Gap: Situated Spatial Language a Case-Study in Connecting Perception and Language

John D. Kelleher

Technological University Dublin, john.d.kelleher@tudublin.ie

Follow this and additional works at: https://arrow.tudublin.ie/aircoth

Part of the Artificial Intelligence and Robotics Commons, Cognition and Perception Commons, and the Cognitive Psychology Commons

Recommended Citation

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License
Situated language is spoken from a particular point of view within a simulated or physical context that is shared with an interlocutor. From theoretical linguistic and cognitive perspectives, situated dialog systems are interesting as they provide ideal testbeds for investigating the interaction between language and perception, at the same time there are a growing number of practical applications, for example robotic systems, where spoken interfaces, capable of situated dialog, promise many advantages (Kelleher, 2003). An open challenge in this domain is the creation of computational models that appropriately ground the semantics of spatial terms within the shared perceptual context. This is partly because of the diversity of factors that impinge on spatial term semantics, including geometry, world knowledge (including functional roles and object dynamics), and human perception.

Many computational models of spatial semantics are based on the concept of a spatial template (Logan and Sadler, 1996). This standard model has been extended in a number of ways. For example, to include frame of reference ambiguity (Kelleher and Costello, 2005; Kelleher and van Genabith, 2006; Dobnik et al., 2014); the impact of distractor objects within the scene (Kelleher and Kruijff, 2005; Costello and Kelleher, 2006; Kelleher and Costello, 2009); and to include the role of human attention and visual perceptual factors in spatial reference resolution (Kelleher et al., 2005; Kelleher, 2006; Regier and Carlson, 2001; Kelleher et al., 2010). At the same time, other research has used corpus based analytics to explore the functional and geometric semantics of prepositions in visually situated spatial reference (Dobnik and Kelleher, 2014; Dobnik et al., 2018). However, to-date relatively little work has been focused on developing an integrated model that accommodates all of these factors.

In recent years, however, deep learning approaches have made significant breakthroughs in a number of areas. An exciting aspect of deep learning is the concept of representation learning from data. In particular, learning the projection of naturally discrete information (e.g. words) into continuous representations (e.g. word embeddings), and also learning vector based inter/multi-modal representations, such as those used in automatic image captioning systems. A number of shortcomings with current deep learning architectures have been identified with respect to their application to spatial language (Kelleher and Dobnik, 2017). However, adopting a modular mechanistic approach to training deep networks may offer a solution to these challenges (Dobnik and Kelleher, 2017).

In light of this, in this paper will review the literature on computational models of spatial semantics and the potential of deep learning models as an useful approach to this challenge.

References


