Neat and Scruffy: On Computational Generation and Interpretation of Spatial Descriptions*

Simon Dobnik

Department of Philosophy, Linguistics & Theory of Science (FLoV)
Centre for Linguistic Theory & Studies in Probability (CLASP)
University of Gothenburg, Sweden
simon.dobnik@gu.se

Physical sciences have developed ways in which space can be described with a high degree of accuracy, for example by measuring distances and angles in coordinate systems. Such measures can be represented on a continuous scale of real numbers and their mathematical modelling and computation is well understood.

However, humans refer to space quite differently. Descriptions such as "the chair is to the left of the table", 'the flowers are in a vase" or "'turn right at the next crossroad" refer to discrete units such as points, regions and volumes. They require common sense knowledge how objects related by a preposition interact with each other. Their semantics take into account aspects of linguistic interaction such as communicative intents of speakers and their conversational partners, for example in negotiation of spatial perspective or frame of reference. Mechanisms of attention are used to select information from different contexts and evaluate potential distractors which makes their interpretation notoriously vague.

Spatial descriptions connect both human conceptual and perceptual domains and therefore, I argue, can only be modelled with computational architectures that combine aspects of neat and scruffy models. The majority of current models consider only the geometric perceptual context as a meaning component of spatial descriptions. We argue that common-sense functional knowledge about object interactions (semantic information about their affordances) and reference to objects in different linguistic interactive contexts can be captured by distributional semantic models commonly used in natural language processing, that is from word co-occurrences in contexts. Contextual distributional semantic models or word embeddings can be trained with deep neural networks alongside with other modalities such image and geometric features in the form of grounded language models. Different families of neural networks can be stacked together as modules and the neural architecture naturally supports information fusion. Bottom-up learning of semantics can be combined with top-down engineering in terms of model design, features and injection of conceptual information from ontologies.

Although grounded language models have proven to be very successful in generation and interpretation of spatial language in tasks such as image captioning and visual questing answering, the problem of spatial cognition and inference is by no means solved. I will discuss findings of our studies what such models learn. I will argue that the majority of the model shortcomings come from the fact that we train them in the scenarios where they have to match patterns rather than model inference, from the neural architecture designs which fail to cover all aspects of spatial semantics and from the biases in training datasets which frequently lead to hallucinations, cases where the perceptual modality is ignored. As spatial cognition is not yet fully understood, questions such as what features are to be modelled, what kind of representations should be used and at what granularity present interesting future challenges for collaborative work between theoretical, experimental and computational research of spatial cognition.

Acknowledgement The research was supported by a grant from the Swedish Research Council (VR project 2014–39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.