A Theory of Content

Mark Steedman, University of Edinburgh

The talk compares collocation-based and extension-based distributional semantics for NLP question answering with respect to compatibility with logical operators. Recent work with Mike Lewis and others seeking to define a novel form of semantics for relational terms using semi-supervised machine learning methods over unlabeled text is described. True paraphrases are represented by the same cluster identifier. Common-sense inference as represented by an entailment graph is represented directly in the lexicon, rather than delegated to meaning postulates and theorem-proving. The method can be applied cross-linguistically, in support of machine translation. Ongoing work extends the method to extract multi-word items, light verb constructions and an aspect-based semantics for temporal/causal entailment. This representation of content has interesting implications concerning the nature of the hidden language-independent conceptual language that must underlie all natural languages in order for them to be learnable by children, but which has so far proved resistant to discovery.