

A THEORY OF CONTENT FOR NLP

Mark Steedman (steedman@inf.ed.ac.uk)

University of Edinburgh, UK

Linguists and computational linguists have come up with some quite useful theories of the semantics of function words and the corresponding logical elements such as generalized quantifiers and negation (Woods 1968; Montague, 1973; Steedman 2012). There has been much less progress in defining a usable semantics for content words.

The effects of this deficiency are very bad: linguists find themselves in the embarrassing position of saying that the meaning of „seek“ is seek’. Computation lists find that their wide coverage parsers, which are now fast and robust enough to parse billions of words of web text, have very low precision as question answerers because, while the answers to questions like „Who wrote ‘The Great Gatsby?’“ are out there on the web, they are not stated in the form suggested by the question, „X wrote ‘The Great Gatsby’“, but in some other form that paraphrases or entails the answer, such as „X’s ‘The Great Gatsby’“.

Semantics as we know it is not provided in a form that supports practical inference over the variety of expression we see in real text.

I’ll discuss recent work with Mike Lewis which seeks to define a novel form of semantics for content words using semi-supervised machine learning methods over unlabeled text. True paraphrases are represented by the same semantic constant. Common-sense entailment 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. If I have time, I will discuss the relation of this representation of content to the prelinguistic language of mind that must underlie all natural language semantics, but which has so far proved resistant to discovery.

1. *Lewis and Steedman*, (2013) „Combined Distributional and Logical Semantics“, *Transactions of the Association for Computational Linguistics*, 1, 179–192.
2. *Lewis and Steedman*, (2014) „Combining Formal and Distributional Models of Temporal and Intensional Semantics“, *Proceedings of the ACL Workshop on Semantic Parsing*, Baltimore, 28–32.