INVITED KEYNOTE PRESENTATION

Distributional Semantics and the Lexicon

Eduard Hovy
Information Sciences Institute
University of Southern California
hovy@isi.edu

The lexicons used in computational linguistics systems contain morphological, syntactic, and occasionally also some semantic information (such as definitions, pointers to an ontology, verb frame filler preferences, etc.). But the human cognitive lexicon contains a great deal more, crucially, expectations about how a word tends to combine with others: not just general information-extraction-like patterns, but specific instantial expectations. Such information is very useful when it comes to listening in bad aural conditions and reading texts in which background information is taken for granted; without such specific expectation, one would be hard-pressed (and computers are completely unable) to form coherent and richly connected multi-sentence interpretations.

Over the past few years, NLP work has increasingly treated *topic signature word distributions* (also called 'context vectors', 'topic models', etc.) as a de facto replacement for semantics. Whether the task is wordsense disambiguation, certain forms of textual entailment, information extraction, paraphrase learning, and so on, it turns out to be very useful to consider a word(sense) as being defined by the distribution of word(senses) that regularly accompany it (in the classic words of Firth, "you shall know a word by the company it keeps"). And this is true not only for individual wordsenses, but also for larger units such as *topics*: the product of LDA and similar topic characterization engines is similar.

In this talk I argue for a new kind of semantics, which is being called Distributional Semantics. It combines traditional symbolic logic-based semantics with (computation-based) statistical word distribution information. The core resource is a single lexico-semantic lexicon that can be used for a variety of tasks, provided that it is reformulated accordingly. I show how to define such a semantics, how to build the appropriate lexicon, how to format it, and how to use it for various tasks. The talk pulls together a wide range of related topics, including Pantel-style resources like DIRT, inferences / expectations such as those used in Schank-style expectation-based parsing and expectation-driven NLU, PropBank-style word valence lexical items, and the treatment of negation and modalities. I conclude by arguing that the human cognitive lexicon has to have the same kinds of properties as the Distributional Semantics lexicon, given the ways people *do things with words*.

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), page 1,
Beijing, August 2010