Clueless: Explorations in unsupervised, knowledge-lean extraction of lexical-semantic information

Invited Talk

Lillian Lee
Department of Computer Science, Cornell University
llee@cs.cornell.edu

I will discuss two current projects on automatically extracting certain types of lexical-semantic information in settings wherein we can rely neither on annotations nor existing knowledge resources to provide us with clues. The name of the game in such settings is to find and leverage auxiliary sources of information.

Why is it that if you *know* I'll give a silly talk, it follows that you know I'll give a talk, whereas if you *doubt* I'll give a good talk, it doesn't follow that you doubt I'll give a talk? This pair of examples shows that the word "doubt" exhibits a special but prevalent kind of behavior known as downward entailingness — the licensing of reasoning from supersets to subsets, so to speak, but not vice versa. The first project I'll describe is to identify words that are downward entailing, a task that promises to enhance the performance of systems that engage in textual inference, and one that is quite challenging since it is difficult to characterize these items as a class and no corpus with downward-entailingness annotations exists. We are able to surmount these challenges by utilizing some insights from the linguistics literature regarding the relationship between downward entailing operators and what are known as negative polarity items — words such as "ever" or the idiom "have a clue" that tend to occur only in negative contexts. A cross-linguistic analysis indicates some potentially interesting connections to findings in linguistic typology.

That previous paragraph was quite a mouthful, wasn't it? Wouldn't it be nice if it were written in plain English that was easier to understand? The second project I'll talk about, which has the eventual aim to make it possible to automatically simplify text, aims to learn lexical-level simplifications, such as "work together" for "collaborate". (This represents a complement to prior work, which focused on syntactic transformations, such as passive to active voice.) We exploit edit histories in Simple English Wikipedia for this task. This isn't as simple (ahem) as it might at first seem because Simple English Wikipedia and the usual Wikipedia are far from a perfect parallel corpus and because many edits in Simple Wikipedia do not constitute simplifications. We consider both explicitly modeling different kinds of operations and various types of bootstrapping, including as clues the comments Wikipedians sometimes leave when they edit.

Joint work with Cristian Danescu-Niculescu-Mizil, Bo Pang, and Mark Yatskar.

Proceedings of the Fourteenth Conference on Computational Natural Language Learning, page 55,
Uppsala, Sweden, 15-16 July 2010.