The Lexical Component of Natural Language Processing

George A. Miller
Cognitive Science Laboratory
Princeton University

Abstract

Computational linguistics is generally considered to be the branch of engineering that uses computers to do useful things with linguistic signals, but it can also be viewed as an extended test of computational theories of human cognition; it is this latter perspective that psychologists find most interesting. Language provides a critical test for the hypothesis that physical symbol systems are adequate to perform all human cognitive functions. As yet, no adequate system for natural language processing has approached human levels of performance.

Of the various problems that natural language processing has revealed, polysemy is probably the most frustrating. People deal with polysemy so easily that potential abiguities are overlooked, whereas computers must work hard to do far less well. A linguistic approach generally involves a parser, a lexicon, and some ad hoc rules for using linguistic context to identify the context-appropriate sense. A statistical approach generally involves the use of word co-occurrence statistics to create a semantic hyperspace where each word, regardless of its polysemy, is represented as a single vector. Each approach has strengths and limitations; some combination is often proposed. Various possibilities will be discussed in terms of their psychological plausibility.