How Hard is Syntax?

Mark Liberman
UNIVERSITY OF PENNSYLVANIA

One of the lessons of recent work on "taggers" is this: sometimes NLP problems that seem very hard, if examples are chosen maliciously, are not very hard in typical cases.

This doesn't mean that the hard cases are easy, just that they don't arise all that often. From the practical point of view, this means that systems with tolerable performance are suprisingly easy to build.

Because the ambiguity of typical cases is not so great, crude methods of ambiguity resolution based on simple local counts produce noisy but often useful analyses. In the case of word tagging, a wide variety of approaches seem to converge on roughly similar levels of performance, with success rates better than .95 on a per-word basis.

This leaves open the question of how recalcitrant the remaining problems will be. It also leaves open the issue of what kind of inductive processes are appropriate for finishing the problem off.

In any case, the difference between "typical" cases and "malicious" cases offers a crack into which a variety of techniques can drive a wedge, whether by supervised or unsupervised learning. We can estimate the width of the crack by comparing the dimensionality of the problem viewed categorically with the empirically-estimated entropy.

Although typical sets of lexical categories for English have about a hundred elements in them, and thus potentially represent around six bits of information per word. the actual conditional entropy of lexical category ("tag") given word (in English) does not seem to be very large. Various approaches to empirical measurement of this quantity (to be detailed in the talk) suggest that the actual amount of information in lexical category assignment ("tagging") is at most about half a bit per word.

Notice that this approach to estimating the entropy of tagging is algorithm-neutral, in the sense that we do not make any assumptions about how the tags are to be assigned, but just look at the empirical tag distributions in a corpus. Can we extend this approach to the problem of syntax as a whole? Given that we know a word and its lexical category, how much information is there in its syntactic relationships in a sentence? Can we estimate this quantity independent of any assumptions about parsing methods?

This talk will present a method of solving this problem, and an empirically-derived bound on the per-word entropy of syntactic structures in English text. Unsurprisingly, this measure turns out to be fairly low. Following the same line, approaches are suggested that result in making parsing look as "easy" as tagging has become.