

Emerging Network Structure in a Darwinised Data-Oriented Parser

Data-Oriented Parsing (DOP; Scha 1990, Bod 1992, 1998) is a method for statistical parsing, whereby novel inputs may be analysed by directly exploiting the statistical regularities present in a parsed, labelled training corpus without any abstract representations being generated; rather, training corpus entries are stored as node-labelled trees, from which fragments (subtrees) may be extracted and recombined to produce multiple possible parses of a novel input, for which the most probable parse is approximated by means of a Monte Carlo sample. Darwinised DOP (DDOP) is a novel *unsupervised* DOP parser (see also Bod 2006), which, unlike previous DOP models, operates *incrementally*, which is to say, it gets its training data one sentence at a time instead of storing its entire training corpus right from the beginning; it builds derivations from subtrees extracted from its previous exemplar-base, but backs off to using a randomly-generated subtree when that can't be done. Each time it parses a new input, the output parse is added to its exemplar-base; when subtrees from the exemplar base are used in constructing the output parse, the output parse will contain new copies of them; thus subtrees are Darwinian replicators, and are selected for generalisability. This paper reports on the first experiments with DDOP, but also reports on the online growing network-structure of DDOP exemplar bases in simulations; a Treebank may be thought of as a single graph-theoretic structure if both the tree-internal edges and edges representing the substitutability relation (equivalent to node-labels) are admitted as network components.

References

- Bod, R. (1992). "A Computational Model of Language Performance; Data-Oriented Parsing". *Proceedings COLING-92*, Nantes, France
- Bod, R. (1998), *Beyond Grammar; An Experience-Based Theory of Language*, Stanford, CA: Centre for the Study of Language and Information.
- Bod, R. (2006). "Unsupervised Parsing with U-DOP", paper given at CONLL 2006, New York, NY.
- Scha, R. (1990). "Taaltheorie en Taaltechnologie: Competence en Performance", in Q. de Kort and G. Leerdam (eds.), *Computertoepassingen in de Neerlandistiek*, Almere: Landelijke Vereniging van Neerlandici (LNVN-jaarboek).