

Darwinised Data-Oriented Parsing Subtrees as replicators, DOP as a Genetic Algorithm

Data-Oriented Parsing (DOP; Scha 1990, Bod 1992, 1998) is a powerful statistical parsing method, whereby novel strings are analysed by directly exploiting the statistical regularities present in a treebank without requiring abstract representations to be generated; rather, arbitrary-depth tree-fragments (subtrees) are extracted from the treebank and recombined to produce a Monte-Carlo sample of possible parses for novel strings, by which the Most Probable Parse (MPP) is approximated. More recently, Bod (2006a, b) has extended the approach to unsupervised parsing (UDOP), where, given a corpus of tokenised strings, UDOP uses all subtrees of all possible binary trees over the strings. The model achieves state-of-the-art results, but at a cost of massive computational complexity. Darwinised DOP (DDOP) is a novel unsupervised DOP parser which exploits a hitherto unnoticed feature of DOP. Unlike previous DOP models, DDOP operates incrementally, being fed one string at a time, parsing it, and adding the parse to its training data, rather than loading up the entire training corpus at the start and storing it statically. Like supervised DOP, it builds derivations from subtrees extracted from its previous exemplar-base, but it backs off to using a randomly-generated subtree when a suitable stored one isn't available. When subtrees from stored exemplars are used in constructing the output parse and the output is added to the training data, the training data gains additional copies of all the subtrees used to compose derivations of the output; thus, subtrees are replicators. Moreover, because subtrees that are more highly generalisable are used more often in the derivations of the MPP, there is a selection pressure favouring generalisability. Therefore, incremental DOP may be run as a Genetic Algorithm; by allowing backoff to the use of random subtrees when a corpus subtree cannot be found, DDOP exploits this genetic property to bootstrap unsupervised DOP without the need for the explosive computational complexity of the "all-subtrees" approach. This paper presents the results from the first tests of the DDOP model.

DDOP possesses a number of interesting cognitive properties; it is the most psychologically plausible DOP model yet, being able to bootstrap syntactic patterns from an initially empty exemplar-base, while only requiring a single tree to be added to the exemplar base for each input. It does seem to be the case that human cognitive development in certain cognitive modalities goes through an initial stage of random productions, followed by selection of favoured patterns; in particular, motor babbling (Meltzoff and Moore 1997, Demiris and Dearden 2005) and vocal babbling (Oudeyer 2006), the modelling of which are in fact intended future directions for this work.

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. (2006a). "An All-Subtrees Approach to Unsupervised Parsing", *Proceedings ACL-COLING 2006*, Sydney.

- Bod, R. (2006b). "Unsupervised Parsing with U-DOP", *Proceedings CONLL 2006*, New York, NY.
- Demiris, Y., and A. Dearden (2005), "From motor babbling to hierarchical learning by imitation: a robot developmental pathway", in L. Berthouze, F. Kaplan, H. Kozima, H. Yano, J. Konczak, G. Metta, J. Nadel, G. Sandini, G. Stojanov and C. Balkenius (Eds.), *Proceedings of the Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*. Lund University Cognitive Studies, 123.
- Meltzoff, A.N., and M.K. Moore (1997), "Explaining Facial Imitation: A Theoretical Model". *Early Development and Parenting*, 6:179-192.
- Oudeyer, P. (2006) *Self-Organization and the Evolution of Speech*. Oxford University Press: Oxford.
- 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 (LVVN-jaarboek).