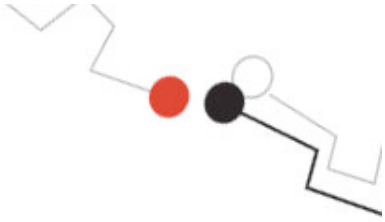
PARIS-
AMSTERDAM
LOGIC MEETINGS
OF YOUNG
RESEARCHERS



Data-Oriented Language Learning - moving beyond negative learnability results
Jelle Zuidema - ILLC
Comments: t.b.a.

Formal modeling of the acquisition of grammar is essential for progress in both linguistic theory and applications in natural language processing. There is an unfortunate tradition in learning theory, however, to focus on the learnability of neat, well-defined classes of formal grammars, and to disregard heuristic methods that perform well on a ragged subset of those grammars but fail a learnability or consistency criterion on the whole set.

I will discuss two examples from quite disparate traditions. The first is the problem of unsupervised learning of a number of (deterministic) formal languages from text in the tradition of Gold (1967). The second is the problem of estimating the weights of a stochastic tree grammar from a labeled tree bank (Bod, 1998; Johnson, 2002). I show that - contrary to received wisdom - negative results in both traditions have little relevance for designing and evaluating language learning algorithms.

The reasons in both cases are in fact exactly the same: only small subsets of the formal classes considered are relevant for the problem of natural language learning. Because the languages we need to learn are languages learned by endless generations before us, the heuristic learning algorithm employed by human learners has in fact defined its own learning problem. Learnability is therefore in some sense guaranteed in "iterated learning" (Kirby & Hurford, 2002; Zuidema, 2003); what we need to worry about instead is whether the learnable class of any proposed algorithm includes the class of natural languages.

These considerations form the motivation for a research program on data-driven, heuristic grammar learning. I will present current work on supervised learning of stochastic tree grammars, and discuss possible extensions to unsupervised learning.

References

Bod, R. (1998). Beyond Grammar: An experience-based theory of language.Stanford, CA: CSLI.
Gold, E. M. (1967). Language identification in the limit. Information and Control (now Information and Computation) 10, 447-474.

Johnson, M. (2002). The DOP estimation method is biased and inconsistent. Computational Linguistics 28, 71-76.
Kirby, S. & Hurford, J. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In: Simulating the Evolution of Language (Cangelosi, A. & Parisi, D., eds.), chap. 6, pp. 121-148. London: Springer Verlag.
Zuidema, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In: Advances in Neural Information Processing Systems 15 (Proceedings of NIPS'02) (Becker, S., Thrun, S. & Obermayer, K., eds.), pp.51-58. Cambridge, MA: MIT Press.