Sampling from probabilistic context-free grammars

Wilker Aziz

March 18, 2015

Probabilistic context-free grammars (PCFGs) are vastly used in natural language processing with applications such as parsing and machine translation. Inference under PCFGs typically requires running a cubic time algorithm (the CKY algorithm) either to compute expectations (necessary in learning) or to make decisions (e.g. finding a good parse). For certain large PCFGs, a complete run of CKY is prohibitively slow. The simplest approximate inference technique involves a pruned run of CKY which discards parses deemed unlikely early on in the search. The problem with arbitrary pruning is that it introduces arbitrary bias harming learning (biased expectations) and inference (local optimum). In this project you will deal with a different class of approximate inference methods, the Monte Carlo (MC) and Markov chain Monte Carlo (MCMC) methods. In these methods, we simulate our target distribution and reason over a reduced statistical sample of the space of solutions.

Exact sampling from a CKY chart is trivial and has been formalised by Chappelier and Rajman (2000). In this project, a strong student will investigate alternative methods which do not require building a complete CKY chart. More specifically, the student will investigate at least one of the following: a Gibbs sampler inspired by the work of Blunsom et al. (2009); a slice sampler previously introduced by Blunsom and Cohn (2010); a novel importance sampler (Neal, 2001).

References

Blunsom, P. and Cohn, T. (2010). Inducing synchronous grammars with slice sampling. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 238–241, Los Angeles, California. Association for Computational Linguistics.

Blunsom, P., Cohn, T., Dyer, C., and Osborne, M. (2009). A Gibbs sampler for phrasal synchronous grammar induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 782–790, Suntec, Singapore. Association for Computational Linguistics.

Chappelier, J.-C. and Rajman, M. (2000). Monte-Carlo sampling for NP-hard maximization problems in the framework of weighted parsing. In Christodoulakis, D., editor, *Natural Language Processing — NLP 2000*, volume 1835 of *Lecture Notes in Computer Science*, pages 106–117. Springer Berlin Heidelberg.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, 11(2):125–139.