Probabilistic Modelling with Context Free Grammars

Wilker Aziz

March 1, 2017

1 Description

Probabilistic context-free grammars (PCFGs) are vastly used in natural language processing with applications such as syntactic parsing, morphological segmentation, topic modelling, and machine translation. PCFGs can be seen as a form of finite-mixture models somewhat similar to first-order hidden Markov models (HMMs). Often this strong independence assumption renders PCFGs unrealistic generative models of data. PCFGs can be extended in several different ways including lexicalisation (Collins, 2003), Markovization (Klein and Manning, 2003), nonterminal refinement (Petrov et al., 2006), to name a few. In this project we are interested in nonparametric extensions. Nonparametric models are capable of adjusting their capacity according to the data. Liang et al. (2007) propose to use a hierarchical Dirichlet process (HDP) in order to induce a PCFG with an unbounded number of nonterminals, Johnson et al. (2006) propose a nonparametric model where the set of rules of the grammar is unbounded and can be extended with rooted sub-trees. Nonparametric extensions to formalisms beyond context-free grammars also exist, for example Cohn et al. (2010) propose a nonparametric tree-substitution grammar, a formalism where rules are (potentially incomplete) sub-trees.

This project will focus on the infinite PCFG (Liang et al., 2007), or HDP-PCFG.

2 Organisation

The project will take place in weeks 10-13 of 2017 (March 6 - March 31) and is worth 6EC (168 hours).

Week 1 A review of important algorithms for PCFGs.

- parsing as deduction (Shieber et al., 1995; Klein and Manning, 2001, 2005) (Kallmeyer, 2010, Chapter 3);
- Earley parsing (Earley, 1970);

- Inside-Outside; (Lari and Young, 1990; Goodman, 1999)
- basic sampling algorithms: ancestral sampling (Chappelier and Rajman, 2000) and Viterbi.

Progress assessment An informal presentation in Week 2 where the student will discuss:

- a deductive system for Earley parsing that can deal with a PCFG in Chomsky normal form (CNF);
- inside-outside;
- ancestral sampling and Viterbi;

and demonstrate the coded implementations.

Week 2 Nonparametric Bayesian modeling methods.

The student will review various nonparametric models used in NLP with special focus on models based on the Dirichlet Process (DP) and that are relevant to parsing:

- Bayesian finite mixture model
- DP mixture model
- Hierarchical DP (HDP)
- HDP-HMM

Sources are (Frigyik et al., 2010), (Heinrich, 2004), and possibly (Beal et al., 2002), (Teh et al., 2004). This will lead the student towards the HDP-PCFG introduced in (Liang et al., 2007).

Weeks 3-4 HDP-PCFG.

Here the student will reproduce the HDP-PCFG (Liang et al., 2007) up to possible changes with respect to inference techniques (the project supervisor will assess the feasibility of implementing variational inference versus a Metropolis-Hastings sampler).
Progress assessment an informal presentation in Week 3 where the student will give an overview of the model and the chosen inference technique.

3 Prerequisites

Interest in statistical parsing and Bayesian statistics. The ideal student will be comfortable with context-free grammars, dynamic programming and statistics. This project requires good knowledge of a programming language (e.g. Python).

4 Final assessment

At the end of the project (March 31), the student is expected to produce a technical report on the several theoretical frameworks explored (e.g. parsing as deduction, statistical modelling, inference). If time permits, this report will include an experimental analysis, but note that the success of this project does not hinge on empirical research.

References

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). The infinite hidden markov model. In *Machine Learning*, pages 29–245. MIT Press.

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.

Cohn, T., Blunsom, P., and Goldwater, S. (2010). Inducing tree-substitution grammars. *J. Mach. Learn. Res.*, 11:3053–3096.

Collins, M. (2003). Head-driven statistical models for natural language parsing. *Comput. Linguist.*, 29(4):589–637.

Earley, J. (1970). An efficient context-free parsing algorithm. *Commun. ACM*, 13(2):94–102.

Frigyik, B. A., Kapila, A., and Gupta, M. R. (2010). Introduction to the dirichlet distribution and related processes. Technical report.

Goodman, J. (1999). Semiring parsing. *Computational Linguistics*, 25(4):573–605.

Heinrich, G. (2004). Parameter estimation for text analysis. Technical report.

Johnson, M., Griffiths, T. L., and Goldwater, S. (2006). Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 641–648, Cambridge, MA, USA. MIT Press.

Kallmeyer, L. (2010). *Parsing Beyond Context-Free Grammars*. Cognitive Technologies. Springer Heidelberg.

Klein, D. and Manning, C. (2005). Parsing and hypergraphs. In Bunt, H., Carroll, J., and Satta, G., editors, *New Developments in Parsing Technology*, volume 23 of *Text, Speech and Language Technology*, pages 351–372. Springer Netherlands.

Klein, D. and Manning, C. D. (2001). Parsing and hypergraphs. In *Seventh International Workshop on Parsing Technologies (IWPT- 2001)*.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lari, K. and Young, S. J. (1990). The estimation of stochastic context-free grammars using the inside–outside algorithm. *Computer Speech and Language*, 4:35–36.

Liang, P., Petrov, S., Jordan, M., and Klein, D. (2007). The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 688–697, Prague, Czech Republic. Association for Computational Linguistics.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia. Association for Computational Linguistics.

Shieber, S. M., Schabes, Y., and Pereira, F. C. N. (1995). Principles and implementation of deductive parsing. *Journal of Logic Programming*, 24:3–36.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2004). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101.