Grammars and Topic Models

Mark Johnson
Centre for Language Sciences and Dept. of Computing
Macquarie University
Sydney, Australia
Mark.Johnson@MQ.edu.au

1 Abstract

Context-free grammars have been a cornerstone of theoretical computer science and computational linguistics since their inception over half a century ago. Topic models are a newer development in machine learning that play an important role in document analysis and information retrieval. It turns out there is a surprising connection between the two that suggests novel ways of extending both grammars and topic models. After explaining this connection, I go on to describe extensions which identify topical multiword collocations and automatically learn the internal structure of named-entity phrases.

The adaptor grammar framework is a nonparametric extension of probabilistic context-free grammars (Johnson et al., 2007), which was initially intended to allow fast prototyping of models of unsupervised language acquisition (Johnson, 2008), but it has been shown to have applications in text data mining and information retrieval as well (Johnson and Demuth, 2010; Hardisty et al., 2010). We'll see how learning the referents of words (Johnson et al., 2010) and learning the roles of social cues in language acquisition (Johnson et al., 2012) can be viewed as a kind of topic modelling problem that can be reduced to a grammatical inference problem using the techniques described in this talk.

2 About the Speaker

Mark Johnson is a Professor of Language Science (CORE) in the Department of Computing at Macquarie University in Sydney, Australia. He was awarded a BSc (Hons) in 1979 from the University of Sydney, an MA in 1984 from the University of California, San Diego and a PhD in 1987 from Stanford University. He held a postdoctoral fellowship at MIT from 1987 until 1988, and has been a visiting researcher at the University of Stuttgart, the Xerox Research Centre in Grenoble, CSAIL at MIT and the Natural Language group at Microsoft Research. He has worked on a wide range of topics in computational linguistics, but his main research areas are computational models of language acquisition, and parsing and its applications to text and speech processing. He was President of the Association for Computational Linguistics in 2003 and is Vice-President elect of EMNLP, and was a professor from 1989 until 2009 in the Departments of Cognitive and Linguistic Sciences and Computer Science at Brown University.

References

Eric A. Hardisty, Jordan Boyd-Graber, and Philip Resnik. 2010. Modeling perspective using adaptor grammars. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 284–292, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark Johnson and Katherine Demuth. 2010. Unsupervised phonemic Chinese word segmentation using Adaptor Grammars. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 528–536, Beijing, China, August. Coling 2010 Organizing Committee.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 641–648. MIT Press, Cambridge, MA.

Mark Johnson, Katherine Demuth, Michael Frank, and Bevan Jones. 2010. Synergies in learning words and their referents. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1018–1026.

Mark Johnson, Katherine Demuth, and Michael Frank. 2012. Exploiting social information in grounded language learning via grammatical reduction. In

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 883–891, Jeju Island, Korea, July. Association for Computational Linguistics.

Mark Johnson. 2008. Using Adaptor Grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, pages 398–406, Columbus, Ohio. Association for Computational Linguistics.