# Book Reviews

## Bayesian Analysis in Natural Language Processing

**Shay Cohen**
(University of Edinburgh)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 35), 2016, xxvii+246 pp; paperback, ISBN 9781627058735, $85.00; ebook, ISBN 9781627054218, $68.00; doi:10.2200/S00719ED1V01Y201605HLT035

*Reviewed by*
*Kevin Duh*
*Johns Hopkins University*



Bayesian techniques are useful tools for modeling a wide range of data and phenomena. Natural language is no exception. The Bayesian approach works as follows:

1. Model the data $x$ probabilistically with $p(x|\theta)$, where $\theta$ are some unknown parameters. For example, this could be a generative story for a sentence $x$, based on some unknown context-free grammar parameters $\theta$.
2. Represent the uncertainty about $\theta$ by a prior distribution $p(\theta)$.
3. Given data, apply Bayes theorem $p(\theta|x) \propto p(x|\theta)p(\theta)$ to find the posterior distribution for the quantities of interest.

This approach enables an elegant and unified way to incorporate prior knowledge and manage uncertainty over parameters. It can also be used to provide capacity control for complex models as an alternative to smoothing. There have been many successful applications of Bayesian techniques in natural language processing (NLP). Some examples include: word segmentation (Goldwater et al. 2009), syntax (Johnson et al. 2007), morphology (Snyder & Barzilay 2008), coreference resolution (Haghighi & Klein 2007), and machine translation (Blunsom et al. 2009).

Cohen's book provides an accessible yet in-depth introduction to Bayesian techniques. It is aimed at a researcher or student who is already familiar with statistical modeling in natural language (i.e., at the level of introductory books such as Manning & Schütze [1999], Jurafsky & Martin [2009]). The stated goal of the book is to "cover the methods and algorithms that are needed to fluently read Bayesian learning papers in NLP and to do research in the area." I believe Cohen successfully achieves this goal, striking a nice balance between breadth and depth of material.

Chapter 1 is a brief review of probability and statistics. It covers prerequisite concepts such as independence, conditional independence, and exchangeability of random variables. The differences between Bayesian and frequentist philosophies are discussed, albeit briefly. In general, the book maintains a pragmatic approach, focusing more on the mathematics and less on the philosophy.

doi:10.1162/COLI_r_00310

Chapter 2 motivates Bayesian techniques in NLP with two distinct examples: latent Dirichlet allocation (LDA) for unsupervised topic modeling and Bayesian linear regression for supervised text analytics. Although Bayesian techniques are most often used for unsupervised problems in NLP (e.g., for addressing problems where the Expectation-Maximization [EM] algorithm fails to find good solutions), the supervised example demonstrates their broader applicability. One aspect that I appreciate about Cohen's exposition is that he strives to distinguish between what is technically feasible with Bayesian techniques vs. what is frequently used in research. This helps avoid potential misconceptions.

Chapter 3 describes the priors $p(\theta)$ that are common in NLP: for example, the Dirichlet distribution, the logistic normal distribution, and non-informative priors such as the Jeffreys prior.

Chapters 4–6 explain Bayesian inference in detail and form the core of the book. How does one combine the data likelihood and the prior to compute the posterior distribution $p(\theta|x) \propto p(x|\theta)p(\theta)$? This can be a computationally difficult problem. Chapter 4 discusses the maximum a posterior (MAP) estimation of $p(\theta|x)$. Chapter 5 covers sampling methods, particularly Gibbs sampling, Metropolis-Hastings sampling, and slice sampling. Chapter 6 describes variational inference, the mean-field approximation, and the variational EM algorithm. In each chapter, algorithm variants that are popular in NLP research are discussed in detail. An example is blocked Gibbs sampling using dynamic programming.

Chapter 7 focuses on Bayesian nonparameterics. These models allow for an infinite-dimensional parameter space, but the actual number of parameters grows with sample size. This is an active field of research. Cohen does a laudable job of explaining the basic mathematics behind the Dirichlet process, which generalizes the Dirichlet distribution. He describes the stick-breaking and Chinese Restaurant Process viewpoints, and shows how the Dirichlet process can be used as a prior in a nonparametric mixture model and a hierarchical topic model.

Chapter 8, the final chapter, demonstrates how Bayesian techniques can be applied to grammar models in NLP. It focuses on parametric and non-parameteric Bayesian models for probabilistic context-free grammars. I wished there was an additional chapter on other applications besides grammar models. However, I think a reader who completes this book would have gained the technical background to do such a survey by themselves.

In summary, Cohen's *Bayesian Analysis in Natural Language Processing* is a good starting point for a researcher or a student who wishes to learn more about Bayesian techniques. It covers the necessary and sufficient knowledge needed to understand papers in this area, and leaves the remaining details as references. It can be viewed as a concise introduction that complements the many excellent statistics textbooks on Bayesian techniques, which tend to be more detailed.

I wish I had this book when I first started to learn Bayesian techniques back in 2009. I recall struggling through several advanced statistics textbooks, before being rescued by Kevin Knight's classic workbook, "Bayesian Inference with Tears."[1] Cohen's book might have saved me some Kleenex.

1 https://www.isi.edu/natural-language/people/bayes-with-tears.pdf.

## References

Blunsom, Phil, Trevor Cohn, Chris Dyer, and Miles Osborne. 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.

Goldwater, Sharon, Thomas Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Haghighi, Aria and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 848–855, Prague.

Johnson, Mark, Thomas L Griffiths, and Sharon Goldwater. 2007. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems*, pages 641–648, Vancouver.

Jurafsky, Daniel and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall.

Manning, Chris and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Snyder, Benjamin and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745, Columbus, OH.

*Kevin Duh* is a senior research scientist at the Human Language Technology Center of Excellence (HLTCOE) and an assistant research professor at the Department of Computer Science, Johns Hopkins University, Baltimore, MD. His research interests lie at the intersection of Natural Language Processing and Machine Learning, particularly on areas relating to machine translation, semantics, and deep learning. Duh's e-mail address is `kevinduh@cs.jhu.edu`.