Linguistic Structure Prediction

Noah A. Smith
Carnegie Mellon University

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 13), 2011, xx+248 pp; paperbound, ISBN 978-1-60845-405-1, $60.00; ebook, ISBN 978-1-60845-406-8, $30.00 or by subscription

Reviewed by
Chris Quirk
Microsoft Research

Noah Smith's ambitious new monograph, *Linguistic Structure Prediction*, "aims to bridge the gap between natural language processing and machine learning." Given that current natural language processing (NLP) research makes heavy demands on machine-learning techniques, and a sizeable fraction of modern machine learning (ML) research focuses on structure prediction, this is clearly a timely and important topic. To address the gaps and overlaps between these two large and well-developed fields in five brief chapters is a difficult feat. The text, though not without its flaws, does an admirable job of building this bridge.

An introductory first chapter surveys current research areas in statistical NLP, cataloging and defining many common linguistic structure prediction tasks. Machine learning students new to the area are likely to find this helpful albeit a bit terse; NLP students will likely consider this section primarily a review. The subsequent chapters change character abruptly, delving into mathematical details and heavy formalism.

Chapter 2 introduces the concept of decoding, presenting five distinct viewpoints on the search for the highest scoring structure. The reader is quickly ushered through graphical models, polytopes, grammars, hypergraphs, and weighted deduction systems, with descriptions based on an example in sequence tagging. The broad coverage, multi-viewpoint discussion encourages the reader to make connections between many distinct approaches, and provides solid formalism for reasoning about decoding problems. It is a comprehensive introduction to the most common and effective decoding approaches, with one significant exception: the recent advances in dual decomposition and Lagrangian relaxation methods. Timing is likely the culprit. This book was developed mainly from 2006 to 2009, whereas dual decomposition did not attain notoriety in our community until a few years later (Rush et al. 2010). Relaxation approaches, though potentially a passing phase, have successfully broadened the reach of simpler decoding techniques into more complicated domains such as structured event extraction. They would have made a nice addition. Regardless, this second chapter equips the reader with sufficient machinery to solve a number of structured prediction problems.

Chapter 3 applies the machinery described in the prior chapter to the problem of supervised structure induction. Probabilistic generative and conditional models are introduced in some detail, followed by a discussion of margin-based methods. Hidden Markov models (HMMs) and probabilistic context-free grammars are introduced in detail, followed by solid descriptions of maximum likelihood estimation and smoothing. The section on conditional models is well written and crucial, because so many commonly used tasks can be treated as sequence modeling using techniques such as conditional random fields. Much of the subject matter introduced abstractly in Chapter 2 is presented in this chapter using specific algorithms. For instance, sequence

modeling is discussed broadly in Chapter 2; the specific algorithms for HMMs are fully defined in Chapter 3. This coarse-to-fine introduction of material may challenge readers who are accustomed to more practical descriptions of material. Were I to teach a course based on this book, I would be tempted to present the third chapter before the second.

Chapter 4 focuses on semisupervised, unsupervised, and hidden variable learning. With a good mix of theory and practical examples, Expectation-Maximization (EM) is introduced and grounded in several problems, then generalized with log-linear models and approximated with contrastive estimation. Hard EM is mentioned in the context of several examples, though a more detailed description of this potentially important technique (cf. Spitkovsky et al., 2010) would bridge the material of Chapters 2 and 4. The chapter then describes Bayesian approaches to NLP, working from theory into specific techniques and landing in models. Finally, a brief section is devoted to the related area of hidden variable learning.

Chapter 5 begins by describing the partition function, as well as inference techniques for the partition function and decoding methods. I found it strange that this important section was postponed so late in the book; much of the material was forward-referenced throughout Chapter 4. Regardless, the techniques are described in a unifying, generic manner. The book concludes with a discussion of minimum Bayes risk decoding, and a few other variants.

Four appendices are devoted to optimization, experimental techniques, maximum entropy, and locally normalized conditional models. All of these sections provide some useful background. The section on hypothesis testing in Appendix B would be especially useful to students new to the area. It can be difficult to pick the correct hypothesis testing method in general, and this problem is exacerbated in structure prediction. This material serves as a good guide for a researcher hoping to evaluate how effective these methods are.

I have some concerns about the intended audience. Descriptions quickly descend into heavy notation and require knowledge of a broad range of mathematical concepts, from marginal polytopes to semirings. I suspect the average NLP graduate student would find it difficult to approach much of the material without a series of courses in probability, statistics, and machine learning. The book is also very theoretical: Few concrete algorithms are provided. Instead, the concepts are introduced using only mathematics and formalism. For readers already conversant in the mapping from mathematical descriptions into concrete algorithms and implementations, this will not be a significant barrier. From the other direction, the structures used in NLP (e.g., dependency trees) are relatively well motivated, though a machine-learning researcher new to the area might benefit from a fuller introduction to NLP. However, the text serves as an effective guide for introducing the machine-learning community into the NLP community, but I feel it would be challenging to use in the other direction.

This may be a personal bias, but I was surprised by the avoidance of machine-translation–related techniques, despite obvious influences. Why resort to the term "decoding" if not because of decipherment and translation? One of the most effective uses of hidden variable learning is in word alignment; it seems like a personal example. Of course, building an effective machine-translation system requires a huge amount of engineering in addition to the underlying theory, but I felt some discussion of the problem and effective techniques would be pertinent.

Despite my struggles with the book's organization and a few important omissions, I must admit that I want a copy for my bookshelf. The author covers a huge amount of material in just under 200 pages, touching on some of the most important algorithmic techniques and viewpoints in modern statistical NLP. At times the text reads like a

survey, touching very briefly on a huge range of topics. Yet the survey is comprehensive and enlightening, tying together a broad range of topics and viewpoints. Younger graduate students may require serious effort to comprehend the full text, but the modern NLP researcher looking to advance the state of the art in structured prediction must truly understand the concepts presented here.

References

Rush, Alexander M., David Sontag, Michael Collins, and Tommi Jaakkola. 2010. On dual decomposition and linear programming relaxations for natural language processing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1–11. Cambridge, MA.

Spitkovsky, Valentin I., Hiyan Alshawi, Daniel Jurafsky, and Christopher D. Manning. 2010. Viterbi training improves unsupervised dependency parsing. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 9–17. Uppsala.

This book review was edited by Pierre Isabelle.

Chris Quirk is a Researcher in the Natural Language Processing group at Microsoft Research. His research interests include machine translation and paraphrase, particularly at the confluence of large data and linguistic analysis tools. Quirk's e-mail address is chrisq@microsoft.com.