

# A Review of *Statistical Language Learning*

Marti A. Hearst

■ *Statistical Language Learning*, Eugene Charniak, The MIT Press, Cambridge, Massachusetts, 1993, 170 pp., ISBN 0-262-03216-3.

**S***tatistical Language Learning* is an introduction to the rapidly burgeoning field of statistical computational linguistics. Several factors have led to the increase in interest in this field, which is heavily influenced by techniques from speech processing. One major factor is the recent availability of large online text collections. Another is a disillusionment with traditional AI-based approaches to parsing and natural language processing (NLP). Charniak is recognized as a distinguished contributor to what he calls traditional AI NLP, which is why it is all the more significant that in the Preface, when speaking of his recent transition to the statistical approach, he writes

... few, if any, consider the traditional study of language from an artificial-intelligence point of view a "hot" area of research. A great deal of work is still done on specific NLP problems, from grammatical issues to stylistic considerations, but for me at least it is increasingly hard to believe that it will shed light on broader problems, since it has steadfastly refused to do so in the past. (p. xvii)

There is a real need for a textbook to serve newcomers to the field of statistical NLP. With this book, Charniak attempts to provide a summary of the basic mathematical tools and algorithms employed by current corpus-based language-analysis research.

Chapter 1 describes standard or traditional approaches to NLP, focusing

on syntactic chart parsing and briefly mentioning semantic processing. This description is too brief for beginners and too simple for those who know the field; its purpose is to situate the ideas from standard NLP that are revisited later in the book in terms of their probabilistic counterparts.

Chapter 2 describes a small fragment of probability and information theory, including brief coverage of probability theory (conditional probability, Bayes law) and entropy, cross-entropy, and Markov chains. Entropy is explained lucidly by way of an appeal to coding theory, and cross-entropy is explained in terms of its role in the evaluation of statistical models.

Chapter 3 describes hidden Markov models and their application to two problems: the estimation of trigram models of language and stochastic part-of-speech assignment. Chapter 4 describes the standard algorithms for training and using hidden Markov models (the Viterbi algorithm for efficient decoding and the forward-backward, or Baum-Welch, algorithm for training). Rather than presenting the proof for the convergence of the training algorithm, Charniak illustrates the behavior of one cycle of training using a detailed example. He also describes the problems that can arise with training (local maxima, overfitting, and critical points).

Chapters 5 and 6 discuss probabilistic parsing using context-free grammars. Chapter 5 shows the advantages of a probabilistic context-free grammar parser over a nonprobabilistic one and touches on the role of probabilistic context-free grammars in grammar induction. Chapter 6 describes the standard training algorithm for assigning probabilities to

context-free grammars: the inside-outside algorithm. Charniak draws parallels between training hidden Markov models to recognize regular languages and training using the inside-outside algorithm to recognize context-free languages.

Chapter 7 focuses on the automatic acquisition of probabilistic context-free grammars, showing how training on existing text collections can help circumvent the problem of a lack of negative examples. Charniak points out the problems inherent in this approach—for example, one data set can yield an infinite number of grammar rules—and describes approaches to constraining the possible induced grammars by restricting the form that the induced grammar rules might take.

Chapters 8, 9, and 10 describe recent research on more isolated aspects of parsing and language analysis. Chapter 8 describes a few approaches to specialized structural ambiguity resolution, focusing on Hindle and Rooth's (1993) lexically based prepositional phrase-attachment training algorithm and two approaches to reducing its parameter space using manually assigned semantic tags. Chapter 9 describes some of the recent research in automated assignment of lexical items into semantically related classes, using cooccurrence information extracted from corpora. Chapter 10 mentions a few approaches to the related subject of corpus-based approaches to word-sense disambiguation.

There is no concluding or summarizing chapter.

Several themes run through the book that can be recognized by practitioners in the field, for example, the problem of sparse data resulting from models with too many parameters and the importance of lexical information for parsing as opposed to strictly syntactic information.

This book has several strong points. As Bill Gale points out on the cover flap, Charniak does a good job of presenting the ideas behind the approaches and is careful to point out their limitations. Furthermore, the expositions of the basics in Chapters

2 to 4 are clear and helpful. Descriptions of the more standard algorithms, such as those for training and using hidden Markov models and the inside-outside algorithm, are also clear and are demonstrated with useful examples.

However, the book is flawed in several ways, some more damaging than others. Charniak states explicitly that he has made no attempt to be comprehensive in his coverage of the material, and of course, opinions will differ over what should be included in a short introduction. However, at times, the omissions are misleading; as one example, no mention is made of the EM (expectation-maximization) algorithm (Dempster, Laird, and Rubin 1977), which subsumes the forward-backward and inside-outside algorithms and is perhaps more useful for those readers who would like to devise their own algorithms. As another example, the coverage of grammar-induction algorithms is too narrow given the book's title; many important approaches predating Charniak's work are entirely absent. As a final example, issues surrounding smoothing do not receive a proper treatment. (As a side point, I found the title somewhat misleading because in my world view at least, language learning is a much broader topic than grammar induction.)

Another serious flaw is a scholarly one: a lack of proper attribution and external references. Although Charniak attributes the forward-backward algorithm to Baum (Baum et al. 1970), there is no reference at all for the inside-outside algorithm (Baker 1979; Lari and Young 1990). This lack is especially distressing because this algorithm does not have the established status of, say, chart parsing (which does receive a reference in the form of a pointer to the survey in Winograd's [1983] syntax book. As another example, although the discussion of entropy is clear, it is also abbreviated, and Charniak presents no pointers whatsoever to further information (one possibility would have been Cover and Thomas [1991]). No general statistics texts are recommended to supplement the material, and the coverage of basic probability

theory is less than adequate.

This book is meant to be used as a textbook, and the exercises seem to assume that many of those who read about the work should be able to implement the algorithms. However, the level of algorithmic description could be more detailed at times. There are tricky aspects to coding hidden Markov models and probabilistic parsers that the unsuspecting reader is likely to encounter. An obvious example is the problem of underflow in the forward-backward algorithm (the repeated multiplication of probability estimates quickly leads to small floating-point numbers). On the surface, this algorithm looks straightforward to implement, but the underflow problem must be faced in any real implementation. Charniak does not even point out that most implementations use the logarithms of the probabilities. Kai Fu Lee's (1989) book, for example, is helpful at describing solutions to various implementation problems.

Finally, the book suffers stylistically, as if the author could not decide what kind of tone to adopt. The prose drifts in tone from that of a textbook to that of an informal lecture to that of a technical paper, sometimes often within the space of one or two pages. The unevenness of the prose at times extends to the organization of the presentation of the material as well. One example is found in the discussion of ergodicity. In the midst of a formal exposition, Charniak writes

Formally, this holds if our language  $L$  is *ergodic*, a point we return to later. (p 34)

One page later, he writes

The reason this works, of course, is that we were guaranteed in advance that the test suite of 100 examples was exactly indicative of the probabilistic model.... Nevertheless it is not hard to make approximations to such a suite and use equation 2.22 with the understanding that the results of the model testing might be off because of biases in the suite. By the way, requiring that the language be "ergodic" is simply a fancy way to say that any sample of the language, if

made long enough, is such a perfect sample. (p. 35ff)

If ergodicity is to be mentioned at all, it should be in the form of a definition when first introduced.

Despite its flaws, this book fills a vacuum in the NLP community because it is the first textbook geared toward this audience that covers the background material and standard algorithms for statistical language analysis. For someone just entering the field or someone teaching an advanced undergraduate or graduate course on the topic, this book is a serviceable introduction if supplemented with a bibliography and other readings as long as the reader is aware that only a narrow view of the field is presented. It most likely will not and should not be the only attempt at an introductory textbook in this area.

### Bibliography

- Baker, J. 1979. Trainable Grammars for Speech Recognition. In Proceedings of the Spring Conference of the Acoustical Society of America, 547-550. Boston, Mass.: Acoustical Society.
- Baum, L. E.; Petrie, T.; Soules, G.; and Weiss, N. 1970. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions in Markov Chains. *The Annals of Mathematical Statistics* 41(1): 164-171.
- Cover, T. M., and Thomas, J. A. 1991. *Elements of Information Theory*. New York: Wiley.
- Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Series B* 34:1-38.
- Hindle, D., and Rooth, M. 1993. Structural Ambiguity and Lexical Relations. *Computational Linguistics* 19(1): 103-120.
- Lari, K., and Young, S. J. 1990. The Estimation of Stochastic Context-Free Grammar Using the Inside-Outside Algorithm. *Computer Speech and Language* 4:35-56.
- Lee, K.-F. 1989. *Automatic Speech Recognition: The Development of the SPHINX System*. Boston, Mass.: Kluwer Academic.
- Winograd, T. 1983. *Language as a Cognitive Process, Volume 1: Syntax*. Reading, Mass.: Addison-Wesley.
- Marti A. Hearst** received her Ph.D. in computer science from the University of California at Berkeley in 1994. She is now a member of the research staff at Xerox Palo Alto Research Center.