Foundations of Statistical Natural Language Processing

Christopher D. Manning and Hinrich Schütze
(Stanford University and Xerox PARC)

Cambridge, MA: The MIT Press, 1999,
xxxvii + 680 pp. Hardbound, ISBN
0-262-13360-1, $60.00

Reviewed by
Lillian Lee
Cornell University

In 1993, Eugene Charniak published a slim volume entitled *Statistical Language Learning*. At the time, empirical techniques to natural language processing were on the rise — in that year, *Computational Linguistics* published a special issue on such methods — and Charniak's text was the first to treat the emerging field.

Nowadays, the revolution has become the establishment; for instance, in 1998, nearly half the papers in *Computational Linguistics* concerned empirical methods (Hirschberg, 1998). Indeed, Christopher Manning and Hinrich Schütze's new, by-no-means slim textbook on statistical NLP — strangely, the first since Charniak's[1] — begins, "The need for a thorough textbook for Statistical Natural Language Processing hardly needs to be argued for". Indubitably so; the question is, is this it?

Foundations of Statistical Natural Language Processing (henceforth FSNLP) is certainly ambitious in scope. True to its name, it contains a great deal of preparatory material, including: gentle introductions to probability and information theory; a chapter on linguistic concepts; and (a most welcome addition) discussion of the nitty-gritty of doing empirical work, ranging from lists of available corpora to in-depth discussion of the critical issue of smoothing. Scattered throughout are also topics fundamental to doing good experimental work in general, such as hypothesis testing, cross-validation, and baselines. Along with these preliminaries, FSNLP covers traditional tools of the trade: Markov models, probabilistic grammars, supervised and unsupervised classification, and the vector-space model. Finally, several chapters are devoted to specific problems, among them lexicon acquisition, word sense disambiguation, parsing, machine translation, and information retrieval.[2] (The companion website contains further useful material, including links to programs and a list of errata.)

In short, this is a Big Book[3], and this fact alone already confers some benefits. For the researcher, FSNLP offers the convenience of one-stop shopping: at present, there is no other NLP reference in which standard empirical techniques, statistical tables, definitions of linguistics terms, and elements of information retrieval appear together; furthermore, the text also summarizes and critiques many individual research papers. Similarly, someone teaching a course on statistical NLP will appreciate the large number of topics FSNLP covers, allowing the tailoring of a syllabus to individual interests. And for those entering the field, the book records "folklore" knowledge that is typically acquired only by word of mouth

1 In the interim, the second edition of Allen's book (1995) did include some material on probabilistic methods, and much of Jelinek's *Statistical Methods for Speech Recognition* (1997) concerns language processing. Also, the forthcoming *Speech and Language Processing* (Jurafsky and Martin, in press) promises to cover many empirical methods.

2 The grouping of topics in this paragraph, while convenient, does not correspond to the order of presentation in the book. Indeed, the way in which one thinks about a subject need not be the organization that is best for teaching it, a point to which we will return later.

3 For the record: 3 lb., 10.7 oz.

or bitter experience, such as techniques for coping with computational underflow. The abundance of numerical examples and pointers to related references will also be of use.

Of course, encyclopedias cover many subjects, too; a good text not only contains information, but arranges it in an edifying way. In organizing the book, the authors have "decided against attempting to present Statistical NLP as homogeneous in terms of mathematical tools and theories" (pg. xxx), asserting that a unified theory, though desirable, does not currently exist. As a result, instead of the ternary structure implied by the third paragraph above — background, theory, applications — fundamentals appear on a need-to-know basis. For example, the key concept of separating training and test data (failure to do so being regarded in the community as a "cardinal sin" (pg. 206)) appears as a subsection of the chapter on n-gram language modeling. It is therefore imperative that the "Road Map" section (pg. xxxv) be read carefully.

This design decision enables the authors to place attractive yet accessible topics early in the book. For instance, word sense disambiguation, a problem students seem to find quite intuitive, is presented a full two chapters before hidden Markov models, even though HMM's are considered a basic technology in statistical NLP. Two benefits accrue to those who are developing courses: students not only receive a more gentle (and, arguably, appetizing) introduction to the field, but can start course projects earlier, which instructors will recognize as a nontrivial point.

However, the lack of an underlying set of principles driving the presentation has the unfortunate consequence of obscuring some important connections. For example, classification is not treated in a unified way: Chapter 7 introduces two supervised classification algorithms, but several popular and important techniques, including decision trees and k-nearest-neighbor, are deferred until Chapter 16. Although both chapters include cross-references, the text's organization blocks detailed analysis of these algorithms as a whole; for instance, the results of Mooney's (1996) comparison experiments simply cannot be discussed. Clustering (unsupervised classification) undergoes the same disjointed treatment, appearing both in Chapter 7 and 14.

On a related note, the level of mathematical detail fluctuates in certain places. In general, the book tends to present helpful calculations; however, some derivations that would provide crucial motivation and clarification have been omitted. A salient example is (the several versions of) the EM algorithm, a general technique for parameter estimation which manifests itself, in different guises, in many areas of statistical NLP. The book's suppression of computational steps in its presentations, combined with some unfortunate typographical errors, risks leaving the reader with neither the ability nor the confidence to develop EM formulations in his or her own work.

Finally, if FSNLP had been organized around a set of theories, it could have been more focused. In part, this is because it could have been more selective in its choice of research paper summaries. Of the many recent publications covered, some are surely, sadly, not destined to make a substantive impact on the field. The book also occasionally exhibits excessive reluctance to extract principles. One example of this reticence is its treatment of the work of Chelba and Jelinek (1998); although the text hails this paper as "the first clear demonstration of a probabilistic parser outperforming a trigram model" (pg. 457), it does not discuss what features of the algorithm lead to its superior results.

Implicit in all these comments is the belief that a mathematical foundation for statistical natural language processing can exist and will eventually develop. The authors, as cited above, maintain that this is not currently the case, and they might well be right. But in considering the contents of FSNLP, one senses that perhaps already there is a thinner book, similar to the current volume but with the background-theory-applications structure mentioned above, struggling to get out.

I cannot help but remember, in concluding, that I once read a review that said something like the following: "I know you're going to see this movie. It doesn't matter what my review says. I could write *my hair is on fire* and you wouldn't notice because you're already out buying tickets". It seems likely that the same situation exists now; there is, currently, no other comprehensive reference for statistical NLP. Luckily, this big book takes its responsibilities seriously, and the authors are to be commended for their efforts.

But it is worthwhile to remember that there are uses for both Big Books and Little Books. One of my

colleagues, a computational chemist with a background in statistical physics, recently became interested in applying methods from statistical NLP to protein modeling.[4] In particular, we briefly discussed the notion of using probabilistic context-free grammars for modeling long-distance dependencies. Intrigued, he asked for a reference; he wanted a source that would compactly introduce fundamental principles that he could adapt to his application. I gave him Charniak (1993).

References

Allen, James. 1995. *Natural Language Understanding*. Benjamin Cummings, second edition.

Charniak, Eugene. 1993. *Statistical Language Learning*. MIT Press.

Chelba, Ciprian and Frederick Jelinek. 1998. Exploiting syntactic structure for language modeling. In *ACL 36/COLING 17*, pages 225–231.

Hirschberg, Julia. 1998. "Every time I fire a linguist, my performance goes up," and other myths of the statistical natural language processing revolution. Invited talk, Fifteenth National Conference on Artificial Intelligence (AAAI-98).

Jelinek, Frederick. 1997. *Statistical Methods for Speech Recognition*. MIT Press.

Jurafsky, Daniel and James Martin. In press. *Speech and Language Processing*. Prentice Hall.

Mooney, Raymond J. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Conference on Empirical Methods in Natural Language Processing*, pages 82–91.

Lillian Lee is an assistant professor in the Computer Science Department at Cornell University. Together with John Lafferty, she has led two AAAI tutorials on statistical methods in natural language processing. She received the Stephen and Marilyn Miles Excellence in Teaching Award in 1999 from Cornell's College of Engineering. Lee's address is: Department of Computer Science, 4130 Upson Hall, Cornell University, Ithaca, NY 14853-7501; e-mail: llee@cs.cornell.edu.

4 Incidentally, FSNLP's commenting on bioinformatics that "As linguists, we find it a little hard to take seriously problems over an alphabet of four symbols" (pg. 340) is akin to snubbing computer science because it only deals with zeros and ones.