Foundations of Statistical Natural Language Processing

by Christopher D. Manning and Hinrich Schütze

MIT Press, 1999
620 pages, hardcover, list price $70.00
ISBN 0-262-13360-1

Review by:
Gerhard Weikum, University of the Saarland, Saarbrücken, Germany
weikum@cs.uni-sb.de

It is a pleasure to write this review of an excellent textbook. Pedagogically this is a real gem, and it is a great source of teaching material for a variety of courses. I discovered the book when I was looking for textbook material that I could use in my class on information retrieval (IR), which is an advanced undergraduate course at my university. Being myself a "core database person", I committed myself to teaching the IR class with the intention to learn a new subject and the expectation that there should be plenty of good textbooks to choose from. I easily found a few decent books, but none of them seemed adequate as the main literature for my class (which is also a matter of personal taste, however). I ended up taking material from different sources, textbooks as well as conference and journal papers (the latter especially for Web search), and the one source to which I referred most and which both my students and myself appreciated most was the book by Manning and Schütze.

The book is about natural language processing (NLP), covering the entire spectrum from parsing and disambiguation, sentence tagging, and machine translation, all the way to text analysis, information extraction, and document retrieval. So the book's main topic is not really IR, but among the 15 chapters of the 680-pages book there is plenty of material that fits right in the mainstream of an IR course. In addition, I discovered a substantial fraction of the core NLP material to be extremely insightful and relevant for text mining and other aspects of IR in a broader sense. In particular, I realized the value of the foundational topics that relate to information theory and statistical learning. Of course, I had been aware of the connection between IR and NLP, but there is a big difference between abstract knowledge and really seeing the relationships in detail when preparing a class. The book was a real eye-opener to me in this regard.

The book is organized in four parts: preliminaries, words, grammar, and applications and techniques. The first part, preliminaries, provides general motivation, linguistic basics, and mathematical foundations from elementary probability and information theory including such widely useful concepts as the Kullback-Leibler divergence. Part two, words, discusses different kinds of word occurrence statistics, the problem of word sense ambiguity and techniques for disambiguation, and a suite of lexical analysis techniques including basics of IR such as vector space similarity measures. This part also introduces various

statistical techniques such as maximum likelihood estimators and hypothesis tests. Part three, grammar, is the mathematically most demanding part, discussing Hidden Markov Models (HMMs) for automatic tagging and techniques for probabilistic parsing of natural language sentences. Finally, the fourth and last part, applications and techniques, covers statistical machine translation techniques, text clustering methods like K-means and EM, classic IR topics like vector space retrieval and LSI, and supervised methods for text classification including Naive Bayes, kNN, and maximum entropy methods.

Overall, the book provides an excellent coverage of material that I found appropriate for an advanced undergraduate course on IR. In addition, it contains sufficient pointers to original research publications for those who want to include even more advanced material in their classes, and, of course, also for curious students. In my own class, I included the book's chapters on linguistic basics, vector space model and LSI, clustering, classification, and HMMs for text segmentation, and I also adopted a large fraction of the foundational material on information theory and statistics (from various chapters). I did not follow the book's sequencing of the material, however; after all, my course was on IR and not on NLP. But I did find it very easy to integrate building blocks from the book into my own course organization and combine these with material from other sources.

The book contains a fair amount of mathematics, and I consider this both appropriate for the topic and one of the book's most valuable assets. None of the theoretical underpinnings require truly advanced math, and the book is fully self-contained. Moreover, the book nicely illustrates all (mathematical) concepts by examples, without compromising the rigor of the presentation. The authors, Manning and Schütze, generally provide excellent guidance to their readers with lots of motivating and illustrating examples; the writing style is very clear and precise. This is one of the very best computer science books that I have seen in the last few years, and I highly recommend it to both teachers and students.