What are the uses of this book? Those who design and implement IR systems and wish to know more about algorithms and coding will find it valuable. And certainly those interested in compression of text and images will find it desirable to have as well. Those who teach IR courses will also find the book useful, especially those who teach about IR algorithms and coding.

In all, this is a well-written book that describes how to build IR systems, with a strong focus on compression methods. But its coverage of general IR should also lead others to take a look at it as well.

References

Baeza-Yates R and Ribeiro-Neto B (1999), Eds. Modern Information Retrieval, McGraw-Hill, New York.
Frakes W and Baeza-Yates R (1992), Eds. Information Retrieval: Data Structures and Algorithms, Prentice-Hall, Englewood Cliffs, NJ.
Hersh W (1996) Information Retrieval: A Health Care Perspective, Springer-Verlag, New York.
Lancaster F and Warner A (1993) Information Retrieval Today, Information Resources Press, Arlington, VA.
Meadow C (1992) Text Information Retrieval Systems, Academic Press, San Diego.

William Hersh
Associate Professor & Chief
Division of Medical Informatics and Outcomes Research
Oregon Health Sciences University
Portland, OR, USA

Foundations of Statistical Natural Language Processing. Christopher D. Manning and Hinrich Schütze. Cambridge, MA: MIT Press; 1999; 620 pp. Price: $60.00 (ISBN: 0-262-13360-1.)

Natural language processing has long seemed to be the magic bullet that will bring information retrieval much closer to human capabilities. It is rather frustrating that 20 years of work along these lines has not produced the best information retrieval systems. Generally speaking, systems based entirely on natural language concepts are not at all competitive with systems based on statistical analysis of texts. In addition, although adding natural language features appears to improve the performance of poor systems, no one has yet shown a way to add these features to the best systems and generate any further improvement. There is a growing suspicion that in fact the river is currently flowing the other way, and that ways of thinking about text that have been developed for the purposes of information retrieval have more to contribute to the problem of natural language processing than vice-versa.

The existence of the large and rapidly growing body of material on application of methods of numerical computation to the analysis of natural language texts is the motivation for this excellent book. The book is intended to serve as a reference manual for researchers, supplemented by a 54-page bibliography, and as a textbook for advanced students in computer science.

In spite of the direction of influence mentioned above, the authors are open minded and point out, for example, that a non-quantitative tagger developed at the University of

Helsinki, called English Constraint Grammar performs better than Markov Model Taggers to which it has been compared.

In essence the statistical approach to natural language processing, as discussed in this book, takes the order of terms in a document to be of paramount importance. (This is in contrast to the "bag of words" approach, in which frequencies of term occurrence or co-occurrence are all that remains of the original documents). It seeks to capture, in mathematical terms, the kind of contextual awareness that human readers use to resolve ambiguity. Methods for doing this are based on one central idea: that the probability that a particular token (word) represents a particular part of speech or particular concept can be modeled using a Markov process. The process implied is the one which humans use for the sequential generation of natural language utterances.

The probabilities of the candidate part of speech tags for any particular token will depend on the tokens that appear before (and after) it in the text. They will also depend on the tags that have been tentatively assigned to the preceding terms in the text. The problem of finding the most probable chain of assignments is then attacked using Viterbi's algorithm, which is well known in Engineering and signal processing.

The book is a gold mine of information about various approaches to natural language processing, and actually presents something of a challenge to anyone who would use it as a textbook. If the instructor is well versed in statistical methods, the challenge is to fill in all of the details of the non-quantitative approach as sketched in this book. Correspondingly, an instructor well-versed in non-quantitative methods will be challenged to fill in the details of the statistical principles as they are sketched in this book. However, given the complete absence of any comparable encyclopedic reference/text book, this book can be recommended without hesitation to a researcher seeking to acquire a deeper understanding of contemporary quantitative methods for natural language processing, or to an instructor seeking to enlarge the group of students who are prepared to make meaningful contributions to one of the most pressing problems facing computational science today. The last two chapters focus on specific issues that are close to information retrieval, including language models, and text categorization.

As befits a contemporary book in the field of information retrieval, this work has a website, available through the MIT Press home site, which is updated from time to time (in October 2000 the date of last update was listed as April 2000). This is a very rich source of links to tools and other natural language resources. At the same time, it faces the problem of any web resource, in that links (particularly those to commercial sites) are in some cases no longer valid. However, the central value is in the book rather than the site, and this book can be strongly recommended to anyone interested in the field.

Paul Kantor
Professor, School of Communication, Information and Library Studies
Rutgers University
New Brunswick, NJ
kantor@scils.rutgers.edu