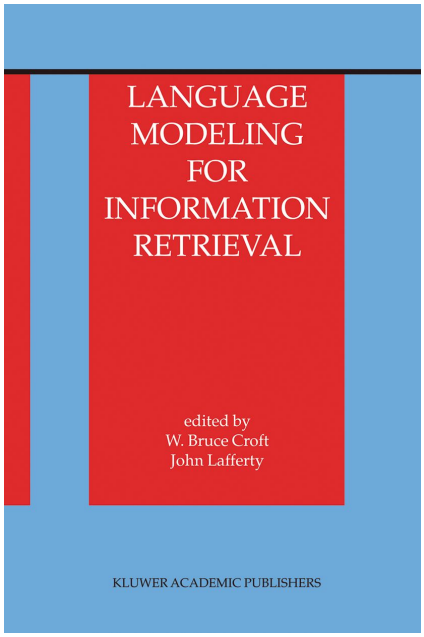

2003, XIV, 246 p.

Printed book

Hardcover

ISBN 978-1-4020-1216-7

- **139,99 € | £119.99**
- ***149,79 € (D) | 153,99 € (A) | CHF 165.50**

B. **Croft**, J. **Lafferty** (Eds.)

Language Modeling for Information Retrieval

Series: The Information Retrieval Series, Vol. 13

A statisticallanguage model, or more simply a language model, is a prob abilistic mechanism for generating text. Such adefinition is general enough to include an endless variety of schemes. However, a distinction should be made between generative models, which can in principle be used to synthesize artificial text, and discriminative techniques to classify text into predefined cat egories. The first statisticallanguage modeler was Claude Shannon. In exploring the application of his newly founded theory of information to human language, Shannon considered language as a statistical source, and measured how weH simple n-gram models predicted or, equivalently, compressed natural text. To do this, he estimated the entropy of English through experiments with human subjects, and also estimated the cross-entropy of the n-gram models on natural 1 text. The ability of language models to be quantitatively evaluated in tbis way is one of their important virtues. Of course, estimating the true entropy of language is an elusive goal, aiming at many moving targets, since language is so varied and evolves so quickly. Yet fifty years after Shannon's study, language models remain, by all measures, far from the Shannon entropy liInit in terms of their predictive power. However, tbis has not kept them from being useful for a variety of text processing tasks, and moreover can be viewed as encouragement that there is still great room for improvement in statisticallanguage modeling.