Language Modeling for Information Retrieval

W. Bruce Croft and John Lafferty (editors)
(University of Massachusetts, Amherst, and Carnegie Mellon University)

Dordrecht: Kluwer Academic
Publishers (Kluwer international series
on information retrieval, edited by W.
Bruce Croft), 2003, xiii+245 pp;
hardbound, ISBN 1-4020-1216-0, $97.00,
£62.00, €99.00

Reviewed by
Paul Thompson
Dartmouth College

The papers in this edited volume are expanded versions of papers originally written for the Workshop on Language Modeling and Information Retrieval held May 31–June 1, 2001, at Carnegie Mellon University in Pittsburgh, Pennsylvania. As the editors note, these papers provide a cross-section of current work on the language-modeling approach to information retrieval, which has become a very active area of research in the past few years. The editors place the papers in this volume in three broad categories: (1) papers addressing the mathematical formulation and interpretation of the language-modeling approach to information retrieval; (2) papers concerned with using the language-modeling approach for ad hoc information retrieval; and (3) papers using the language-modeling approach for other information retrieval application areas, including topic tracking, classification, and summarization. This book provides an excellent introduction to this new field of research by bringing together extended versions of papers from many of the field's leading researchers.

Stated simply, the goal of the language-modeling approach to information retrieval is to predict the probability that the language model of a particular document being considered for retrieval could have generated the user's query. By casting the document retrieval problem in this way, language-modeling techniques that have been developed over many years, particularly for speech recognition, can be applied to document retrieval. On the other hand, at least in this simple statement, the language-modeling approach appears to ignore the concept of relevance. By contrast, the traditional approach to probabilistic information retrieval, as expressed, for example, in the probability-ranking principle, explicitly states that the goal of probabilistic information retrieval is to predict the probability that a document will be judged relevant by a user, taking into account all evidence available to the retrieval system, which then ranks the documents in decreasing order of probability of relevance. As mentioned by the editors in the preface, this concern with the relationship of the language-modeling approach to relevance was one of the underlying themes of the workshop, and the issue is taken up by some of the authors.

The three chapters addressing the mathematical formulation and interpretation of language modeling for information retrieval are Lafferty and Zhai's "Probabilistic Relevance Models Based on Document and Query Generation," Lavrenko and Croft's "Relevance Models in Information Retrieval," and Sparck Jones et al.'s "Language Modeling and Relevance." Taken together, these three chapters illustrate the controversies with respect to the theoretical status of the language-modeling approach. Lafferty and Zhai claim to show an equivalence between the underlying probabilistic semantics

of the language-modeling approach and the standard probabilistic model of information retrieval. In particular they disagree with Sparck Jones et al.'s contention that the language-modeling approach, as usually stated, makes sense only if there is a single relevant document that generates a user's query. Sparck Jones et al., rejecting the simple formulation of the language-modeling approach, take pains to informally sketch what a sounder theoretical construction might be like. The result, although incomplete, is complex. It is unclear whether there are any theoretical advantages in abandoning the standard probabilistic model. Lavrenko and Croft also take up the issue of relevance, introducing the concept of a relevance model for information retrieval, that is, a language model reflecting word frequencies in the class of documents relevant to a particular information need. One of their motivations for introducing relevance models is to overcome one of the major disadvantages of the language-modeling framework: its difficulty in incorporating user interaction. They present many experimental results supporting the retrieval effectiveness of their formal model.

Four chapters discuss language modeling in the context of ad hoc information retrieval: Greiff and Morgan's "Contributions of Language Modeling to the Theory and Practice of IR," Xu and Weischedel's "A Probabilistic Approach to Term Translation for Cross-Lingual Retrieval," Manmatha's "Applications of Score Distributions in Information Retrieval," and Zhang and Callan's "An Unbiased Generative Model for Setting Dissemination Thresholds." Greiff and Morgan argue that statistical estimation in language modeling addresses the bias-variance trade-off and that the importance of language-modeling for information retrieval is that it focuses the attention of information retrieval research on this issue. Thus the good experimental results for the language modeling approach reported throughout this book may be due more to its improved statistical estimation techniques than to the use of language modeling as a theoretical framework.

The remaining three chapters describe applications of language modeling to related information retrieval tasks (i.e., topic tracking, text classification, and summarization): Kraaij and Spitters's "Language Models for Topic Tracking," Teahan and Harper's "Using Compression-Based Language Models for Text Categorization," and Mittal and Witbrock's "Language Modeling Experiments in Non-extractive Summarization." These chapters support the book's premise that the language-modeling approach shows promise not only for ad hoc retrieval, but also for other related information retrieval activities.

Information retrieval and computational linguistics have been more or less closely related fields since the 1950s. Statistical approaches have always played an important role in information retrieval. Within the field of computational linguistics, corpus linguistics has emerged over the past 20 years as an increasingly influential approach. Language-modeling techniques originally developed to support speech recognition are now transforming the field of probabilistic information retrieval. However, as alluded to by Sparck Jones et al., the well-established role of language modeling in speech processing and its exploration in machine translation and summarization are activities fundamentally different from document retrieval. It is to be expected that other areas of human-language technology, such as dialogue modeling and mixed-initiative interaction, might also find more application in information retrieval research.

Paul Thompson is a senior research engineer and lecturer at the Thayer School of Engineering at Dartmouth College. His research interests include probabilistic information retrieval and natural language processing. Thompon's address is Institute for Security Technology Studies, Dartmouth College, 45 Lyme Road, Suite 200, Hanover, NH 03755; e-mail: Paul.Thompson@dartmouth.edu.