Language Modeling for Information Retrieval

THE KLUWER INTERNATIONAL SERIES ON INFORMATION RETRIEVAL

Series Editor:
W. Bruce Croft
University of Massachusetts, Amherst

Also in the Series:

MULTIMEDIA INFORMATION RETRIEVAL: ***Content-Based Information Retrieval from Large Text and Audio Databases****, by Peter Schäuble;* ISBN: 0-7923-9899-8

INFORMATION RETRIEVAL SYSTEMS: ***Theory and Implementation,*** *by Gerald Kowalski;* ISBN: 0-7923-9926-9

CROSS-LANGUAGE INFORMATION RETRIEVAL, *edited by Gregory Grefenstette;* ISBN: 0-7923-8122-X

TEXT RETRIEVAL AND FILTERING: ***Analytic Models of Performance****, by Robert M. Losee;* ISBN: 0-7923-8177-7

INFORMATION RETRIEVAL: UNCERTAINTY AND LOGICS: ***Advanced Models for the Representation and Retrieval of Information,*** *by Fabio Crestani, Mounia Lalmas, and Cornelis Joost van Rijsbergen;* ISBN: 0-7923-8302-8

DOCUMENT COMPUTING: ***Technologies for Managing Electronic Document Collections****, by Ross Wilkinson, Timothy Arnold-Moore, Michael Fuller, Ron Sacks-Davis, James Thom, and Justin Zobel;* ISBN: 0-7923-8357-5

AUTOMATIC INDEXING AND ABSTRACTING OF DOCUMENT TEXTS, *by Marie-Francine Moens*; ISBN 0-7923-7793-1

ADVANCES IN INFORMATIONAL RETRIEVAL: ***Recent Research from the Center for Intelligent Information Retrieval****, by W. Bruce Croft*; ISBN 0-7923-7812-1

INFORMATION RETRIEVAL SYSTEMS: ***Theory and Implementation,*** **Second Edition**, *by Gerald J. Kowalski and Mark T. Maybury;* ISBN: 0-7923-7924-1

PERSPECTIVES ON CONTENT-BASED MULTIMEDIA SYSTEMS, *by Jian Kang Wu; Mohan S. Kankanhalli;Joo-Hwee Lim;Dezhong Hong;* ISBN: 0-7923-7944-6

MINING THE WORLD WIDE WEB: ***An Information Search Approach****, by George Chang, Marcus J. Healey, James A. M. McHugh, Jason T. L. Wang;* ISBN: 0-7923-7349-9

INTEGRATED REGION-BASED IMAGE RETRIEVAL, *by James Z. Wang;* ISBN: 0-7923-7350-2

TOPIC DETECTION AND TRACKING: Event-based Information Organization, *edited by James Allan;* ISBN: 0-7923-7664-1

Language Modeling for Information Retrieval

Edited by

W. Bruce Croft
University of Massachusetts, Amherst U.S.A.

and

John Lafferty
Carniege Mellon University, Pittsburgh, U.S.A.

SPRINGER-SCIENCE+BUSINESS MEDIA, B.V.

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-90-481-6263-5 ISBN 978-94-017-0171-6 (eBook)
DOI 10.1007/978-94-017-0171-6

Printed on acid-free paper

Contents

Preface

A statistical language model, or more simply a *language model*, is a probabilistic mechanism for generating text. Such a definition 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 categories.

The first statistical language 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 well 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 text.[1] The ability of language models to be quantitatively evaluated in this 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 limit in terms of their predictive power. However, this 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 statistical language modeling.

In the past several years a new framework for information retrieval has emerged that is based on statistical language modeling. The approach differs from traditional probabilistic approaches in interesting and subtle ways, and is fundamentally different from vector space methods. It is striking that the language modeling approach to information retrieval was not proposed until the late 1990s; however, until recently the information retrieval and language modeling research communities were somewhat isolated. The communities are

[1] C. E. Shannon. Prediction and entropy of printed English. *Bell System Technical Journal*, Vol. 30, pp. 51–64, 1951.

now beginning to work more closely together, and research at a number of sites has confirmed that the language modeling approach is an effective and theoretically attractive probabilistic framework for building IR systems. But there is still groundwork to do in understanding the basics of this new approach, and many possibilities exist for further development of the framework.

This book contains a collection of papers that together give an overview of the recent research activity in the area of language modeling for information retrieval. The book grew out of a workshop on this topic that took place at Carnegie Mellon University on May 31 and June 1, 2001, under the support of ARDA. Several of the presentations were extended into the papers that appear here.

The papers reflect the major themes of the workshop, and can be thought of as falling into three broad, overlapping categories. Several papers address the mathematical formulation and interpretation of the language modeling approach. These papers focus on the important but subtle and elusive concept of relevance, which was a recurring theme of the workshop. Another category is the development of the language modeling approach for improving ad hoc retrieval—the problem of finding documents to answer unstructured queries on unrestricted topics. In these papers the focus is on the problems of variance reduction, better estimation of relevance models, improving cross-language retrieval, score normalization and the distribution of scores in relevant and non-relevant documents. The third category of papers treats the use of language modeling methods in other important application areas within information retrieval, including topic tracking, text classification, and summarization.

The diversity of the papers collected here is an indication of the richness of language modeling methods for approaching information retrieval problems. We believe that the papers provide an interesting cross-section of current work, and hope that they will inspire future research in this rapidly progressing field.

BRUCE CROFT

JOHN LAFFERTY

Contributing Authors

Jamie Callan is an Associate Professor in the School of Computer Science at Carnegie Mellon University. He earned his Ph.D. from the University of Massachusetts, Amherst.

Bruce Croft is a Distinguished Professor in the Department of Computer Science at the University of Massachusetts, Amherst. He received his Ph.D. from the University of Cambridge.

Warren Greiff is a Principal Scientist at The MITRE Corporation. He earned his Ph.D from the University of Massachusetts, Amherst.

David Harper is a Research Professor in the School of Computing at The Robert Gordon University, Aberdeen, Scotland, and Director of the Smart Web Technologies Center.

Wessel Kraaij is a senior researcher and project manager at the department of Multimedia Technology and Statistics of TNO TPD (Netherlands Organization of Applied Scientific Research).

Djoerd Hiemstra wrote his Ph.D. thesis on the use of language models for information retrieval. He is currently an Assistant Professor at the Database Group of the University of Twente.

John Lafferty is an Associate Professor in the Computer Science Department at Carnegie Mellon University. He received his Ph.D. from Princeton University.

Victor Lavrenko is a Research Assistant in the Department of Computer Science at the University of Massachusetts, Amherst.

R. Manmatha is a Research Assistant Professor in the Department of Computer Science at the University of Massachusetts, Amherst. He received his Ph.D. from the University of Massachusetts, Amherst.

Vibhu Mittal is a Senior Scientist at Google corporation and an adjunct faculty member in the School of Computer Science at Carnegie Mellon University. He earned his Ph.D from the University of Southern California.

William Morgan is a Staff Scientist at MITRE Corporation.

Stephen Robertson is a researcher at Microsoft Research Cambridge, and also a Professor at City University, London. For some years, he has contributed to the Okapi system.

Karen Sparck Jones is Professor of Computers and Information at the University of Cambridge, and has worked in language and information processing since the 1950s.

Martjin Spitters is a researcher at the department of Multimedia Technology and Statistics of TNO TPD (Netherlands Organization of Applied Scientific Research).

William Teahan is a lecturer in Computer Science in the School of Informatics at the University of Wales, Bangor. He gained his Ph.D. from the University of Waikato, New Zealand.

Ralph Weischedel is a Principal Scientist at BBN Technologies with a broad range of interests in language processing technologies. He received his Ph.D. from the University of Pennsylvania.

Michael Witbrock is Director, Knowledge Formation & Dialogue, at Cycorp in Austin, TX. He earned his Ph.D in Computer Science from Carnegie Mellon University.

Jinxi Xu is a Scientist at BBN technologies. He earned his Ph.D from the University of Massachusetts, Amherst.

Hugo Zaragoza is a researcher at Microsoft Research Cambridge. He earned his Ph.D. from the University of Paris 6.

ChengXiang Zhai is an Assistant Professor at the University of Illinois at Urbana-Champaign. He received his Ph.D. from Carnegie Mellon University.

Yi Zhang is a Ph.D. candidate in the School of Computer Science at Carnegie Mellon University. She earned her M.S. degree from Carnegie Mellon University and her B.S. degree from Tsinghua University.