# Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval

**Paul S. Jacobs (editor)**
(Research and Development Center, General Electric Company)

Hillsdale, NJ: Lawrence Erlbaum
Associates, 1992, viii + 281 pp.
Hardbound, ISBN 0-8058-1188-5, $59.95;
Paperbound, ISBN 0-8058-1189-3, $27.50

*Reviewed by*
*Stephanie W. Haas*
*University of North Carolina at Chapel Hill*

This book grew out of the 1990 AAAI Symposium of the same name, which explored "new ways to take advantage of the power of on-line text" (p. vii). Nine papers were selected from the symposium, and most of them have been significantly expanded with more discussion or examples or updated with more recent research. Two of the participants, Zernik and Sparck Jones, contributed new papers to this collection. In addition, three papers by Jacobs, Hobbs et al., and Croft and Turtle have been added that provide a context for the other papers. According to Jacobs, "This volume is aimed at an audience of computer professionals who have at least some knowledge of natural language and IR, but it has also been prepared with advanced students in mind" (p. viii). The book is arranged in three sections: "Broad-scale NLP," "'Traditional' information retrieval," and "Emerging applications." The first two sections describe current issues and methods in use in NLP and IR. The third section presents research applications that represent the opportunities made available by combining NLP and IR techniques. Each chapter is followed by its own bibliography, and there is a brief index of names and topics for the book as a whole. The use of the word *intelligent* in the title is a little misleading, and Jacobs admits that some contributors had doubts about it. Rather than referring to AI knowledge–based systems, it is meant to describe systems that are "fast, effective, and helpful" (p. vii).

In the preface, Jacobs comments on the importance of the cross-fertilization of NLP and IR for text retrieval, an area that has received increasing attention in recent years. IR has expanded from its traditional task of producing a set of relevant document citations from a bibliographic database, and now includes diverse activities such as fact retrieval, text categorization, and the construction and use of hypertext links. (See, for example, the chapters by Maarek, Hayes, and Salton and Buckley, respectively.) Most of the papers in this collection describe specific projects or systems; and the details of their components and processes, the liberal use of examples, and the descriptions of their results give the reader a good idea of the sense of excitement felt by researchers and practitioners in this field.

In the first chapter, Jacobs introduces the major themes that recur, implicitly and explicitly, throughout the book. Overall, these could be described as a need for more "realism" in the size and nature of the tasks that research in NLP and IR addresses. Applications must deal with very large corpora (millions of words), must be able to obtain their results in very short periods of time, and must do so with an acceptable level of performance. For example, a story classifier must be able to keep up with a real-time news feed, and a retrieval system must be able to extract relevant documents from an enormous collection before the user gets impatient. These requirements place constraints on the kinds of processing that are practical, and lead to interesting discussions on weak versus strong methods, and the depth of processing and representation that is really necessary for various types of retrieval. The chapters in the first section of the book, especially those by Wilks et al., Hirst and Ryan, and McDonald, focus on these issues. In addition, what happens when these approaches fail is important, since a real-world text processing system cannot usually wait for human intervention and guidance. Given the range of vocabulary, topic, and text style that systems must handle, one must assume that there will be unfamiliar words and structures. Robustness and using incomplete information are discussed in the chapters by Hobbs et al. and McDonald, among others.

As Jacobs notes in his brief introduction to the second section, IR has always had a more immediate concern about evaluation than most NLP research. "If a technique does not significantly increase the percentage of relevant text that a user sees, or significantly decrease the amount of irrelevant text, it is considered altogether unproven" (p. 125). There is a great deal of discussion in IR as to the best way to measure retrieval performance, but DARPA's MUC-3 and TIPSTER evaluation workshops demonstrate the advantages of more rigorous evaluations that allow comparisons between different approaches to be made. These workshops are mentioned briefly in the book; details of MUC-3 are given by Sundheim (1991).

The chapters in the second section clarify the fact that there are really two parts to the question of how to combine NLP and IR: Which NLP techniques are useful in IR? To which stages of the IR process should they be applied? IR could be (very) loosely described as finding relationships between a query and some text, and a variety of approaches are presented, from representing structural relationships between words, phrases, and larger sections of text (Lewis; Salton and Buckley) to inferring relationships from Bayesian networks of concept nodes (Croft and Turtle). Sparck Jones provides a warning that full text approaches should not be applied blindly, and Lewis echoes her in saying that it is important to look at previous IR research, including that using nonautomatic methods, for clues as to which areas are likely to be the most fruitful. Sparck Jones especially recommends a look at indexing as a profitable area. While this selection of papers might not provide a representative introduction to current IR research, it does demonstrate the diversity apparent in the field.

Papers in the third section describe some of the work that falls at the junction of NLP and IR. These range from a system already in commercial use (the categorization system described by Hayes), through a system that has had some limited use and evaluation (Maarek's GURU, which constructs a help system from online documentation), to a highly experimental system for identifying if an agent is "in favor of, neutral, or opposed to the event" described in the text (Hearst). The first chapter, by Stanfill and Waltz, seems a little out of place. It gives a brief description of how statistics and AI might fit into the text retrieval picture. This is a very rich topic, and probably deserves more space to fully develop these issues than it has here.

The overall quality of the papers is good. By and large, they give a clear conceptual background for the research, justifying the importance of the themes presented above. Sufficient details and examples are provided to give a general understanding of the research, and the reader can refer to work listed in the bibliographies for more. The organization of the book works fairly well, although I might quibble with the order of a couple of the chapters (Croft and Turtle and Sparck Jones, for instance). It is unfortunate that some of the few proofreading errors are in the bibliographies, such as the omission of the Robertson and Hancock-Beaulieu reference cited on page 171, and the title missing from the Kirkpatrick et al. reference on page 56. (See the references below.) However, Jacobs does mention that the book was rushed to press in order to keep the contents as current as possible.

The reader who is only slightly acquainted with NLP and IR, or only familiar with one of these areas, will find this an introduction to the issues and (dare I say?) signs of progress that make this such an exciting area of research today. Those with more experience will find a good discussion of both the theoretical bases of the field and the practical concerns of implementing text-based systems. Finally, those who have read the original symposium working notes will find interesting additions to the earlier papers—a chance to find out "what happened next."

**References**

Kirkpatrick, S., Gelatt, C., and Vecchi, M. (1983). "Optimization by simulated annealing." *Science*, **220**, 671–680.

Robertson, S., and Hancock-Beaulieu, M. (1992). "On the evaluation of IR systems." *Information Processing and Management*, **28**(4), 457–466.

Sundheim, Beth, (ed.) (1991). *Proceedings of the Third Message Understanding Conference (MUC-3)*. Morgan Kaufmann.

*Stephanie W. Haas* is an assistant professor in the School of Information and Library Science at the University of North Carolina at Chapel Hill. She received a Ph.D. from the University of Pittsburgh in 1989. Her research interests include linguistic approaches to information retrieval, uses of machine-readable dictionaries, and applications of case grammar. Her address is: School of Information and Library Science, CB# 3360, 100 Manning Hall, University of North Carolina, Chapel Hill, NC 27599-3360; e-mail: stephani@ils.unc.edu