Natural Language Information Retrieval

Tomek Strzalkowski (editor)
(General Electric, Research and Development)

Dordrecht: Kluwer Academic
Publishers (Text, speech and language
technology series, edited by Nancy Ide
and Jean Véronis, volume 7), 1999,
xxv+384 pp; hardbound, ISBN
0-7923-5685-3, $144.00, £84, Dfl 240.00

Reviewed by
Simon Corston-Oliver
Microsoft Research

Information Retrieval (IR) in this collection of 14 original papers is broadly construed to include document retrieval, information extraction, question answering, clustering and classification. In the introduction Strzalkowski asks the provocative question Why hasn't NLP had more success in IR?, the answer to which ought to be of interest to readers of *Computational Linguistics*. Unfortunately, the majority of the papers in the book completely fail to address this question, and several of the papers do not discuss applications of NLP to IR at all. A brief summary of the papers demonstrates the range of topics covered.

Karen Sparck Jones, in "What is the role of NLP in text retrieval?," gives an overview of linguistically motivated indexing (LMI) and nonlinguistic indexing (NLI). LMI is not suitable for queries with few words, yet as more words are added, the conjunction of search terms benefits NLI, raising the bar against which LMI is being compared. Sparck Jones concludes that LMI might still be useful for displaying informative information about documents.

Christian Jacquemin and Evelyne Tzoukermann, in "NLP for term variant extraction: Synergy between morphology, lexicon, and syntax," perform phrase normalization on the basis of full morphological analysis and patterns over parts of speech and syntactic constituents. They provide an overview of finite-state automata for morphological analysis and rule ordering for derivational affixation in French, with a tangential section on Spanish.

Gerda Ruge, in "Combining corpus linguistics and human memory models for automatic term association," draws on psycholinguistic research to improve models of spreading activation within a semantic network sensitive to head/modifier relationships.

Alan F. Smeaton, in "Using NLP or NLP resources for information retrieval tasks," after experiments with matching entire syntactic analyses for TREC yielded results that were much worse than traditional *tf.idf* measures, has experimented with selectively using NLP resources such as WordNet. Retrieval of picture captions with manual word sense disambiguation outperforms a *tf.idf* baseline.

Tomek Strzalkowski, Fang Lin, Jin Wang, and Jose Perez-Carballo, in "Evaluating natural language processing techniques for information retrieval: A TREC perspective," search for appropriate ways to weight linguistic and nonlinguistic representations of document content, and explore expansions of the query based on selecting entire paragraphs. Manually selected paragraphs yield substantial gains in precision.

Jussi Karlgren, in "Stylistic experiments in information retrieval," uses automated genre classification to rerank documents so that documents of the type preferred by human judges are more highly weighted.

Ellen Riloff and Jeffrey Lorenzen, in "Extraction-based text categorization: Generating domain-specific role relationships automatically," describe two advances in information extraction: a method for generating extraction patterns automatically, without the need for a corpus of manual annotations, and a new variation on their augmented relevancy signatures algorithm that does not rely on semantic features.

Yorick Wilks and Robert Gaizauskas, in "LASIE jumps the gate," provide a technical description of the components of GATE, elements of which were used in the LASIE MUC system.

Joe Zhou, in "Phrasal terms in real-world IR applications," describes refinements of algorithms for extracting domain-specific phrases and prototype systems that use these phrases for summarization and categorization.

Paul Thompson and Christopher C. Dozier, in "Name recognition and retrieval performance," apply mature name recognition technology to TREC 1996 queries that contain personal names. They achieve improvements in precision by weighting the names differently from other terms.

Jim Cowie, in "Collage: An NLP toolset to support Boolean retrieval," provides a technical description of a toolkit for automatic analysis of topics in NL form.

Louise Guthrie, Joe Guthrie, and James Leistensnider, in "Document classification and routing: A probabilistic approach," perform statistical document classification based on term occurrence. They augment the classification scores if the documents satisfy a Boolean specification of the words that members of a given class must contain.

Julian M. Kupiec, in "MURAX: Finding and organizing answers from text search," performs conventional information retrieval followed by shallow linguistic analysis within the returned set of documents to answer *Trivial Pursuit* questions using *Grolier's Encyclopedia*. Different strategies are used according to the question type.

Marti A. Hearst, in "The use of categories and clusters for organizing retrieval results," provides an overview of approaches to classification and clustering and the UI problems associated with presenting results to users.

For the most part, the papers are coherent. Many authors have used the space available to them to combine a solid exposition of material that occurs elsewhere in the literature with the presentation of exciting new directions. What this collection lacks is a unifying theme.

As noted above, several of the papers do not address the role of NLP in IR at all. In a few cases, authors acknowledge the rubric of the book only in order to make the point that they are not interested in NLP. Hearst, for example, comments (p. 336), "Although this chapter is part of a book about natural language processing and information retrieval, it does not discuss in detail the natural language processing aspects of categorization and clustering." Similarly, Guthrie, Guthrie, and Leistensnider (p. 290) state:

> Since the theme of this book is the incorporation of the techniques of NLP into the problems associated with information retrieval, the reader might incorrectly assume that in this paper we are attempting to "understand" a document or portions of a document in much the same way as may be necessary for machine translation or for summarization.

Instead, Guthrie, Guthrie, and Leistensnider suggest that mathematical modeling of terms amounts to a coarser-grained representation of meaning that suffices for classification.

Since most papers do not address Strzalkowski's question, it is as well that he suggests a few possible answers of his own: NLP analyses are simply not deep enough, or perhaps the traditional focus on precision and recall causes us to overlook new areas where NLP might be more appropriate. Smeaton proposes a peculiar answer: NLP systems have historically been tailored to machine translation, which he views as an easier task than IR, and are therefore difficult to apply to IR.

The closest thing in this collection to a majority answer to Strzalkowski's question is the view that permeates many of the papers in this volume: NLP is believed to be simply too computationally expensive or too poor in quality to be practical, especially for commercially feasible systems. Zhou, for example, considers even part-of-speech tagging too computationally expensive in the face of enormous textual databases. Even for pure research systems such concerns appear paramount. Kupiec, for example, concedes that more sophisticated NL analysis might be helpful for question answering, but emphasizes that the less sophisticated analyses he uses yield response times of less than four seconds per question.

As we witness the incremental improvements in the field achieved by limiting NL analyses to what is currently tractable, we naturally wonder what the upper bound is for those techniques. What performance might we expect from more expensive, deeper analyses? Fortunately, some researchers are willing to look to the horizon to determine how successful NL techniques might become. Strzalkowski et al. and Smeaton, for example, even perform analysis by hand to determine the peak theoretical performance of their algorithms. Establishing which avenues are worth pursuing enables the focused development of tractable NL analysis techniques.

The lack of a unifying theme and the fact that the papers cover such a wide range of topics mean that the book as a whole is unlikely to appeal to individual researchers. Conversely, all researchers in the field of NLP and IR are likely to find something of interest.

There is one table that is not referred to in the text (p. 143) and typographical errors abound, including occasional missing words (leading to interesting garden path phenomena), mistranscribed phonetic symbols, and transposed acronyms. A well-structured index more than compensates for these minor flaws.

Simon Corston-Oliver is a discourse linguist in the Natural Language Processing Group at Microsoft Research, Redmond, USA. His research interests include functional, quantitative, and machine learning approaches to the study of discourse phenomena, with application to summarization, document retrieval, document macro structure, and the study of aboutness. Corston-Oliver's address is Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA; e-mail: simonco@microsoft.com.