TREC: Experiment and Evaluation in Information Retrieval

Ellen M. Voorhees and Donna K. Harman (editors)
(National Institute of Standards and Technology)

Cambridge, MA: The MIT Press (Digital libraries and electronic publishing series, edited by William Y. Arms), 2005, x+462 pp; hardbound, ISBN 0-262-22073-3, $45.00

Reviewed by
Nicola Stokes
National ICT Australia, University of Melbourne

TREC, the Text Retrieval Conference, is the information retrieval (IR) community's annual evaluation forum, sponsored by the U.S. Department of Defense and the National Institute of Standards and Technology (NIST). The event is split into tracks (e.g., ad hoc retrieval, filtering, question answering) that encapsulate different research agendas in the community. The end result of each track meeting is an overview report written by the track organizers and a collection of technical reports by the track participants. Many of these reports, after some refinement, find their way into leading IR-related conferences such as SIGIR, and every few years a special issue dedicated to a particular TREC or a TREC track is published. The purpose of the present book is fourfold: to collate and distill 12 years' worth of experiments (1991–2003) into a single volume; to provide some historical perspective on the evolution of the tasks; to share some of the general findings across tracks; and to encourage participants to take an introspective look at their progress and ask the question, *What next for TREC?*

Despite TREC's obvious focus on ad hoc retrieval (i.e., given a query return a ranked list of relevant documents), this book has a surprising amount to offer the natural language processing (NLP) community, particularly to researchers interested in question answering (QA) and text summarization, and to a lesser extent researchers concerned with the application of information extraction (IE), machine translation, speech processing, and language-generation technologies. It must be stressed, however, that this is not a book for readers looking for an introduction to IR concepts; there are many adequate textbooks that already fill this need such as that of Baeza-Yates and Ribeiro-Neto (1999). Instead it should be viewed as a starting point for researchers who are using standard IR techniques, such as passage retrieval or a term-weighting function, and would like to investigate the state of the art as determined by TREC's evaluation results. In this review, I will make reference to these NLP interests where appropriate.

The book consists of three parts: The first provides a three-chapter overview of TREC, structured around its different tracks, its evaluation methodology, and its test collections; the second consists of seven chapters on selected track reports; and the final part contains seven reports from the perspective of the participants, many of whom have devoted their efforts to multiple TREC tasks over the years. All of these contributions are of a high quality; this is not surprising given that most of the participants have been working on their respective areas for at least the duration of their TREC track(s). Each chapter is followed by its own bibliography, and a comprehensive 12-page index at the back of the book contains entries for keywords and referenced authors. I came across a few editorial oversights, but nothing that significantly downgrades the quality of this publication.

Chapter 1 is written by the editors, Ellen Voorhees and Donna Harman, who each contribute four chapters to this volume. The main objective of the first chapter is to set

the scene for the book's structure by amalgamating the 21 tracks run over the course of TREC's history into the seven different streams discussed in Part II: ad hoc retrieval; retrieval on the Web; noisy text retrieval; multilingual retrieval; interactive retrieval; routing and filtering; and question answering. Many of the less successful and/or short-running tracks, such as the NLP and genomics tracks, are briefly alluded to in this chapter. Readers who are interested in the word sense disambiguation, conceptual indexing, and thesaurus-based expansion experiments of the mid-nineties are advised to trawl through the participant reports on the TREC Web site (http://trec.nist.gov/). Although it has been shown that synonymy, homonymy, and polysemy do not contribute as much to retrieval performance degradation as was once thought (Krovetz and Croft 1992; Sanderson 2000), it seems that the techniques once used to address these phenomena now have a new role to play in domains dominated by technical language, such as the scientific literature used in the new genomics track (Hersh and Bhupatiraju 2003). IE researchers will also find this task interesting as its secondary goal (in the 2003 track) was the annotation of gene references in Medline records and full scientific articles.

When TREC was first established, one of its primary motivating factors was to validate the *test collection evaluation paradigm* introduced by the Cranfield experiments of the 1960s (Cleverdon, Mills, and Keen 1966). At the core of this experimental methodology was the idea that live users could be removed from the evaluation loop, thus simplifying the evaluation and allowing researchers to run in vitro–style experiments in a laboratory with just their retrieval engine, a set of queries, a test collection, and a set of judgments (i.e., a list of relevant documents). In chapters 2 and 3, Harman, Voorhees, and Buckley trace the history of the standardization of the TREC evaluation methodology, from the development of the test collections and relevance judgments (using methods such as pooling) to the convergence of the evaluation measures to a set of precision-oriented metrics such as mean average precision. Most of this discussion is centered upon the evaluation of ad hoc retrieval systems; however, for TREC tasks that do not return a ranked list of results, the underlying paradigm is still the same, and its extension to tracks such as QA and Filtering is discussed by the respective track organizers in Part II of the book.

The second part of the book offers insights from some of TREC's main contributors. All of these chapters roughly follow the same formula: They begin with a detailed overview of the participant approaches for each year of the track and end with some general comments on the success of the track, its challenges, and future directions. None of this discussion is detailed enough to facilitate the building of a track-specific IR system but each chapter does include a useful bibliography for further study. Chapters that are of particular interest to NLP researchers include Donna Harman's "Beyond English," which discusses how IR techniques that have proved their worth on English text can be easily adapted to address retrieval over other languages. These adaptations include building a language-specific stemming algorithm and stopword list. For languages that are very different from English, such as Chinese, additional preprocessing steps such as word segmentation must be implemented. In fact, this outcome is a general finding for nearly all of the TREC tasks discussed in this book. That is, standard statistical IR methods have proven to be quite capable of performing in a resilient way on different types of data and in many different retrieval scenarios. The one exception to this, however, is the QA track. Researchers have found that as the required answer quality increases (say from relevant passage to exact answer snippet), traditional IR approaches will benefit from the analysis provided by NLP technologies such as named entity classification, coreference resolution, and inference

mechanisms, which facilitate question classification, question reformulation, and semantic pattern matching. Ellen Voorhees's chapter, "Question answering at TREC," chronicles how QA systems and the track evaluation methodology have evolved with the changing QA task definition from 1999 to 2003. Unfortunately, there is no discussion on what remains to be done in this area and how future TRECs can address these concerns.

In this part of the book, NLP researchers may also be interested in the discussion of machine translation applied to multilingual IR, and the effects of noisy text from automatic speech recognition and optical character recognition on the retrieval process. The latter is an area of research that has been largely ignored by the NLP community but may become more relevant when large-scale evaluations such as the Document Understanding Conference turn their attention to summarization over noisy data. Similarly, as NLP applications are deployed and evaluated in a Web environment, the conclusions drawn by Web track participants will become more relevant. For example, future QA system evaluation will more than likely take place within the Web track, in which case researchers will be anxious to explore methods for exploiting Web-specific features, such as link evidence and URL structure, which can be used to improve their passage-level retrieval and answer-extraction processing phases. David Hawking and Nick Craswell's chapter, "The very large collection and Web tracks," is a real gem; their frank discussion of the reluctance of commercial search engines to engage in the track (even after goading) and the difficulty of imposing TREC high-recall–oriented evaluation in a Web environment, when search engine users are more concerned with high-precision returns, is engrossing. It is obvious that many of the issues in the Web track have led to heated debate in the TREC community, and this passion and excitement is captured in this chapter.

Part III of the volume offers commentaries from the individual groups on their personal TREC experiences, their specific contributions, and how TREC has helped to shape their research agendas. The aim of these chapters, it would seem, is to balance out the vague system description offered in the track reports from Part II. The first four chapters are written by long-standing TREC participants who have made significant contributions in terms of research output and the organizational effort that a large evaluation forum like TREC requires. Chapter 11 provides details on the University of Massachusetts INQUERY retrieval system, which uses a Bayesian inference network to combine evidence of relevance from different analysis components. What is interesting about this work is the authors' observation that many of our modern-day NLP tasks can trace their origins back to work that first surfaced in the IR community. For example, named entity recognition and classification (NERC) was being investigated by IR researchers even before the advent of the TREC campaign. Although these attempts at improving the standard bag-of-words document representation led to only moderate gains in performance, NERC is now an important component in recent QA systems. Chapters 12, 13, and 14 also describe systems that had a major impact on the TREC community: the OKAPI system at City University London, the PIRCS system developed at City University of New York, and the SMART system developed at Cornell. In the case of the latter, Cornell carefully stored a version of their system after each ad hoc retrieval track. Subsequent experiments using each of these versions has shown that the TREC initiative helped double retrieval performance in the first eight years of this track. More modest improvements have been seen since then, and many researchers believe that the next leap forward for IR will only happen if the user is reintroduced into the loop. In Stephen Robertson's enjoyable chapter on the OKAPI BM25 term-weighting scheme, he goes as far as to suggest that the sheer success of TREC and its cost-effective

evaluation paradigm may have actually delayed advancements in the area of interactive IR.

The remaining chapters in this part of the book are from the University of Waterloo (Chapter 15), the European Union–funded Twenty-One project (Chapter 16), and IBM (Chapter 17). The latter two chapters will be of interest to NLP researchers. IBM provides a commercial perspective on how TREC has helped improve some of their products and services. This chapter offers the most detailed look at the experiences of a participant in each of the QA tracks and shows how NLP and IR researchers in the same institution can leverage off each others' expertise. The Twenty-One project chapter investigates the use of language modeling in the context of the ad hoc retrieval and Web tracks. It also provides a historical look at the relationship between probabilistic IR and the more recently embraced language-modeling approach. Language-modeling techniques have proved very popular in IR circles in recent years because of the following attractive properties: They provide a theoretical justification for the weights assigned to terms in weighting schemes such as OKAPI BM25 and they are a proven effective retrieval model. In this chapter, the authors also show that language models can provide a clean method for combining evidence of relevance from sources other than document content.

The book finishes with an epilogue, written by Karen Spärck Jones, which looks at the impact TREC has had on retrieval research, and suggests how it should be adapted to address the changing needs of information provision in areas such as the Web and company intranets. This chapter picks up some of the discussion from the Web track chapter in Part II, offering a retort to some difficult questions relating to the relevance of the TREC evaluation paradigm in an era when the commercial success of Web technologies does not rely on the publication of experimental results. One criticism I have of this discussion is that it contains no explicit references to the Web chapter or any of the other participant contributions in this volume, and I cannot help but think that the overall cohesion of the book would have greatly benefited from this.

The most interesting aspect of this chapter is its vision of a single integrated information processing and management system, where a user's search experience is enhanced by summarization, filtering, QA, IE, and translation technologies. To some extent, current search engines already offer these services; however, Spärck Jones points out that "a collection of buttons on a menu isn't proper integration." Instead, future systems should be capable of processing heterogeneous data types and providing a targeted response that selects the appropriate component technology given the user's information need, for example, a ranked list, a summary, a factoid, a relevant passage. Obviously, this ambitious vision will require a cross-fertilization of ideas from the IR and NLP communities. This for me is the strongest motivation for NLP to keep up-to-date with advances in IR and for readers who are interested enough to read this review to consider buying a copy of this book.

Epilogue: Although this book was published in 2005, its contents were written before the TREC 2003 meeting. Since then, many of the tracks under discussion in this volume have stopped or have morphed into new tracks with slightly different agendas. Hence, I think it's fitting that I add a short epilogue of my own here, and say that Spärck Jones's predictions were indeed correct. Five new tracks have been introduced since the writing of this book that tackle retrieval over domains other than news and the Web: company intranets in the Enterprise track; scientific literature in the Genomics track; information-seeking behavior over e-mail (the spam track); and, for the first time this year, the blogosphere (the blog track) and the legal domain (the legal track). TREC Volume II is certain to be equally engaging.

References

Baeza-Yates, Ricardo A. and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval.* Addison-Wesley Longman, Boston.

Cleverdon, C. W., J. Mills, and M. Keen. 1966. Factors determining the performance of indexing systems, vol. 2: Test results. Technical report, Aslib Cranfield Research Project, Cranfield, England.

Hersh, William and R. T. Bhupatiraju. 2003. TREC genomics track overview. Ellen M. Voorhees and L. P. Buckland, editors. *TREC 12, Proceedings of the Twelfth Text Retrieval Conference*, Gaithersburg, Maryland: National Institute of Standards and Technology Special Publication 500-255. *http://medir.ohsu.edu/∼hersh/trec-03-genomics.pdf*.

Krovetz, Robert and W. Bruce Croft. 1992. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141.

Sanderson, Mark. 2000. Retrieving with good sense. *Information Retrieval*, 2(1):49–69.

Nicola Stokes is a Research Fellow at the Victoria Lab of National ICT Australia, University of Melbourne. Her research focuses on the development of robust linguistic analysis techniques (e.g., lexical cohesion analysis, textual entailment and paraphrase identification, and toponym resolution) for use in NLP and IR applications such as ad hoc retrieval, text summarization, question answering, and text classification. Stokes's address is the Department of Computer Science and Software Engineering, ICT Building, 111 Barry Street, Carlton, Victoria 3053, Australia; e-mail: nstokes@csse.unimelb.edu.au.