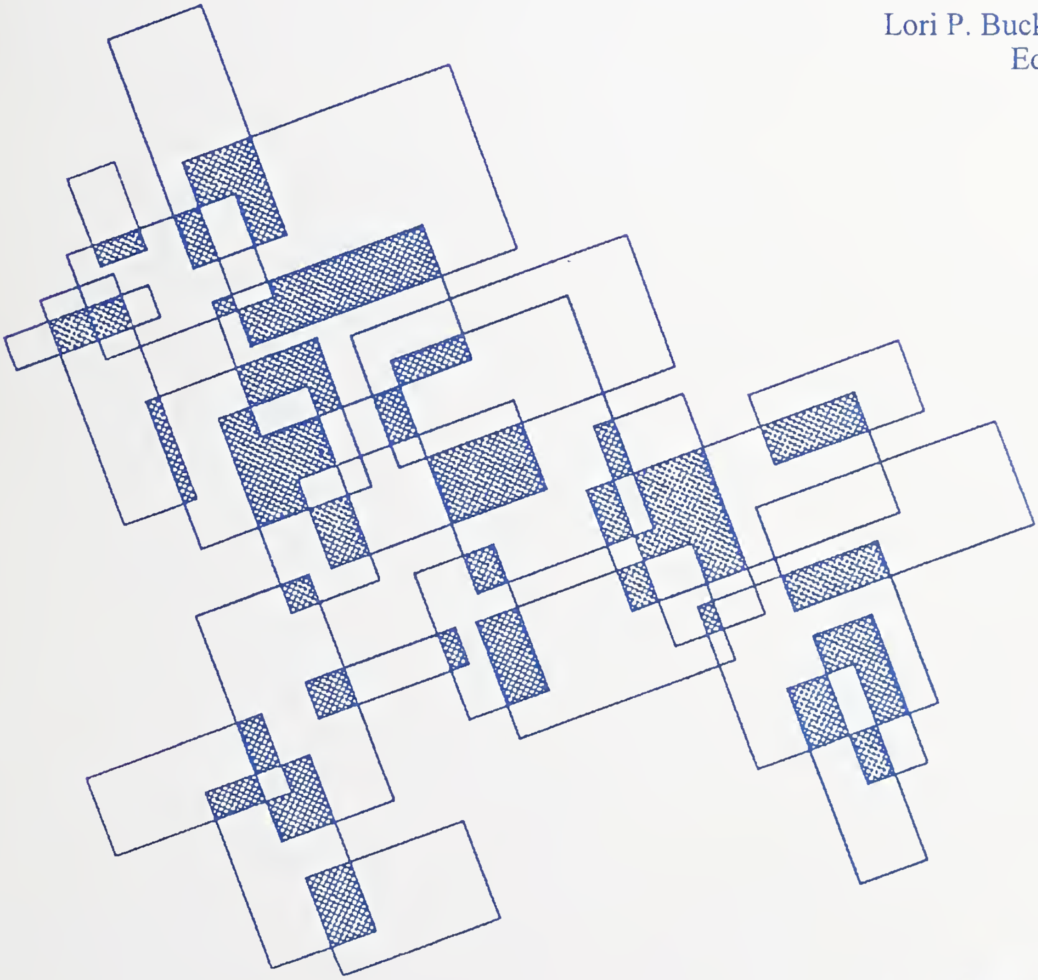
NIST Special Publication 500-261

Information Technology: The Thirteenth Text REtrieval Conference, TREC 2004

Ellen M. Voorhees
and
Lori P. Buckland
Editors



NIST
National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

The National Institute of Standards and Technology was established in 1988 by Congress to "assist industry in the development of technology ... needed to improve product quality, to modernize manufacturing processes, to ensure product reliability ... and to facilitate rapid commercialization ... of products based on new scientific discoveries."

NIST, originally founded as the National Bureau of Standards in 1901, works to strengthen U.S. industry's competitiveness; advance science and engineering; and improve public health, safety, and the environment. One of the agency's basic functions is to develop, maintain, and retain custody of the national standards of measurement, and provide the means and methods for comparing standards used in science, engineering, manufacturing, commerce, industry, and education with the standards adopted or recognized by the Federal Government.

As an agency of the U.S. Commerce Department's Technology Administration, NIST conducts basic and applied research in the physical sciences and engineering, and develops measurement techniques, test methods, standards, and related services. The Institute does generic and precompetitive work on new and advanced technologies. NIST's research facilities are located at Gaithersburg, MD 20899, and at Boulder, CO 80303. Major technical operating units and their principal activities are listed below. For more information visit the NIST Website at http://www.nist.gov, or contact the Publications and Program Inquiries Desk, 301-975-3058.

Office of the Director

- National Quality Program
- International and Academic Affairs

Technology Services

- Standards Services
- Technology Partnerships
- Measurement Services
- Information Services
- Weights and Measures

Advanced Technology Program

- Economic Assessment
- Information Technology and Applications
- Chemistry and Life Sciences
- Electronics and Photonics Technology

Manufacturing Extension Partnership Program

- Regional Programs
- National Programs
- Program Development

Electronics and Electrical Engineering Laboratory

- Microelectronics
- Law Enforcement Standards
- Electricity
- Semiconductor Electronics
- Radio-Frequency Technology[1]
- Electromagnetic Technology[1]
- Optoelectronics[1]
- Magnetic Technology[1]

Materials Science and Engineering Laboratory

- Intelligent Processing of Materials
- Ceramics
- Materials Reliability[1]
- Polymers
- Metallurgy
- NIST Center for Neutron Research

Chemical Science and Technology Laboratory

- Biotechnology
- Process Measurements
- Surface and Microanalysis Science
- Physical and Chemical Properties[2]
- Analytical Chemistry

Physics Laboratory

- Electron and Optical Physics
- Atomic Physics
- Optical Technology
- Ionizing Radiation
- Time and Frequency[1]
- Quantum Physics[1]

Manufacturing Engineering Laboratory

- Precision Engineering
- Manufacturing Metrology
- Intelligent Systems
- Fabrication Technology
- Manufacturing Systems Integration

Building and Fire Research Laboratory

- Applied Economics
- Materials and Construction Research
- Building Environment
- Fire Research

Information Technology Laboratory

- Mathematical and Computational Sciences[2]
- Advanced Network Technologies
- Computer Security
- Information Access
- Convergent Information Systems
- Information Services and Computing
- Software Diagnostics and Conformance Testing
- Statistical Engineering

[1] At Boulder, CO 80303

[2] Some elements at Boulder, CO

NIST Special Publication 500-261

Information Technology: The Thirteenth Text Retrieval Conference, TREC 2004

Ellen M. Voorhees and
Lori P. Buckland
Editors

*Information Technology Laboratory
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940*

August 2005

U.S. Department of Commerce
Carlos M. Gutierrez, Secretary

Technology Administration
Michelle O'Neill, Acting Under Secretary of Commerce for Technology

National Institute of Standards and Technology
William A. Jeffrey, Director

Reports on Information Technology

The Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST) stimulates U.S. economic growth and industrial competitiveness through technical leadership and collaborative research in critical infrastructure technology, including tests, test methods, reference data, and forward-looking standards, to advance the development and productive use of information technology. To overcome barriers to usability, scalability, interoperability, and security in information systems and networks, ITL programs focus on a broad range of networking, security, and advanced information technologies, as well as the mathematical, statistical, and computational sciences. This Special Publication 500-series reports on ITL's research in tests and test methods for information technology, and its collaborative activities with industry, government, and academic organizations.

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

National Institute of Standards and Technology
Special Publication 500-261
Natl. Inst. Stand. Technol.
Spec. Publ. 500-261
132 pages (August 2005)
CODEN: NSPUE2

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON: 2005

For sale by the Superintendent of Documents, U.S. Government Printing Office
Internet: bookstore.gpo.gov—Phone: (202) 512-1800—Fax: (202) 512-2250
Mail: Stop SSOP, Washington, DC 20402-0001

Foreword

This report constitutes the proceedings of the 2004 edition of the Text REtrieval Conference, TREC 2004, held in Gaithersburg, Maryland, November 16–19, 2004. The conference was co-sponsored by the National Institute of Standards and Technology (NIST), the Advanced Research and Development Activity (ARDA), and the Defense Advanced Research Projects Agency (DARPA). Approximately 200 people attended the conference, including representatives from 21 different countries. The conference was the thirteenth in an on-going series of workshops to evaluate new technologies for text retrieval and related information-seeking tasks.

The workshop included plenary sessions, discussion groups, a poster session, and demonstrations. Because the participants in the workshop drew on their personal experiences, they sometimes cite specific vendors and commercial products. The inclusion or omission of a particular company or product implies neither endorsement nor criticism by NIST. Any opinions, findings, and conclusions or recommendations expressed in the individual papers are the authors' own and do not necessarily reflect those of the sponsors.

The sponsorship of the U.S. Department of Defense is gratefully acknowledged, as is the tremendous work of the program committee and the track coordinators.

Ellen Voorhees
August 2, 2005

TREC 2004 Program Committee

Ellen Voorhees, NIST, chair
James Allan, University of Massachusetts at Amherst
Chris Buckley, Sabir Research, Inc.
Gordon Cormack, University of Waterloo
Susan Dumais, Microsoft
Donna Harman, NIST
David Hawking, CSIRO
Bill Hersh, Oregon Health & Science University
David Lewis, Ornarose Inc.
John Prager, IBM
John Prange, U.S. Department of Defense
Steve Robertson, Microsoft
Mark Sanderson, University of Sheffield
Karen Sparck Jones, University of Cambridge, UK
Ross Wilkinson, CSIRO

TREC 2004 Proceedings

Overview Papers

Other Papers

(contents of these papers are found on the TREC 2004 Proceedings CD)

Appendix

(contents of the Appendix are found on the TREC 2004 Proceedings CD)

Papers: Alphabetical by Organization

(contents of these papers are found on the TREC 2004 Proceedings CD)

LexiClone
LexiClone Inc. and NIST TREC

Language Computer Corporation
UB at TREC 13: Genomics Track

Macquarie University
AnswerFinder at TREC 2004

Meiji University
Meiji University Web, Novelty and Genomic Track Experiments

Microsoft Research
Overview of the TREC 2004 Terabyte Track

Microsoft Research Asia
Microsoft Research Asia at Web Track and Terabyte Track of TREC 2004

Microsoft Research Ltd
Microsoft Cambridge at TREC 13: Web and Hard Tracks

MIT Computer Science and Artificial Intelligence Laboratory
Answering Multiple Questions on a Topic From Heterogeneous Resources

MSR Cambridge
TREC 2004 Web Track Experiments at CAS-ICT

National Library of Medicine
Knowledge-Intensive and Statistical Approaches to the Retrieval and Annotation of Genomics MEDLINE Citations

National University of Singapore
Experience of Using SVM for the Triage Task in TREC 2004 Genomics Track

National University of Singapore at the TREC 13 Question Answering Main Task

National Taiwan University
Identifying Relevant Full-Text Articles for GO Annotation Without MeSH Terms

Similarity Computation in Novelty Detection and Biomedical Text Categorization

National Institute of Standards and Technology
Overview of the TREC 2004 Question Answering Track

Overview of the TREC 2004 Robust Track

University of Illinois at Chicago
UIC at TREC 2004: Robust Track

University of Illinois at Urbana-Champaign
UIUC in HARD 2004--Passage Retrieval Using HMMs

The University of Iowa
Novelty, Question Answering and Genomics: The University of Iowa Response

University of Lethbridge
University of Lethbridge's Participation in TREC 2004 QA Track

University of Limerick
Question Answering Using the DLT System at TREC 2004

University of Maryland, Baltimore County
Knowledge-Intensive and Statistical Approaches to the Retrieval and Annotation of Genomics MEDLINE Citations

University of Maryland College Park
Knowledge-Intensive and Statistical Approaches to the Retrieval and Annotation of Genomics MEDLINE Citations

Improving Passage Retrieval Using Interactive Elicition and Statistical Modeling

University of Massachusetts
UMass at TREC 2004: Novelty and HARD

Indri at TREC 2004: Terabyte Track

University of Massachusetts, Amherst
HARD Track Overview in TREC 2004
High Accuracy Retrieval from Documents

The University of Melbourne
Melbourne University 2004: Terabyte and Web Tracks

University of Michigan
The University of Michigan in Novelty 2004

Can We Get A Better Retrieval Function From Machine?

University of North Carolina at Chapel Hill
University of North Carolina's HARD Track Experiments at TREC 2004

University of Waterloo
Overview of the TREC 2004 Terabyte Track

Approaches to High Accuracy Retrieval:
Phrase-Based Search Experiments in the HARD Track

Domain-Specific Synonym Expansion and Validation for Biomedical Information Retrieval
(MultiText Experiments for TREC 2004)

University of Wisconsin, Madison
Exploiting Zone Information, Syntactic Rules, and Informative Terms in Gene Ontology
Annotation of Biomedical Documents

U.S. Department of Defense
Question Answering with QACTIS at TREC 2004

Virginia Tech
Can We Get A Better Retrieval Function From Machine?

York University
York University at TREC 2004: HARD and Genomics Tracks

Papers: Organized by Track

(contents of these papers are found on the TREC 2004 Proceedings CD)

Genomics

Alias-i, Inc.
Phrasal Queries with LingPipe and Lucene: Ad Hoc Genomics Text Retrieval

Aqsaqal Enterprises
DIMACS at the TREC 2004 Genomics Track

Biogen Idec Corporation
TREC 2004 Genomics Track Overview

California State University San Marcos
Categorization of Genomics Text Based on Decision Rules

Cedar/Buffalo
UB at TREC 13: Genomics Track

ConverSpeech LLC
Concept Extraction and Synonymy Management for Biomedical Information Retrieval

David D. Lewis Consulting
DIMACS at the TREC 2004 Genomics Track

Dublin City University
Experiments in Terabyte Searching, Genomic Retrieval and Novelty Detection for TREC 2004

Erasmus Medical Center
MeSH Based Feedback, Concept Recognition and Stacked Classification for Curation Tasks

The German University in Cairo
The GUC Goes to TREC 2004: Using Whole or Partial Documents for Retrieval and Classification in the Genomics Track

Indiana University
WIDIT in TREC 2004 Genomics, Hard, Robust and Web Tracks

Indiana University Bloomington
TREC 2004 Genomics Track Experiments at IUB

Language Computer Corporation
UB at TREC 13: Genomics Track

University of California Berkeley
BioText Team Experiments for the TREC 2004 Genomics Track

University College Dublin
Experiments in Terabyte Searching, Genomic Retrieval and Novelty Detection for TREC 2004

University of Edinburgh
TREC Genomics 2004

University Hospital of Geneva
Report on the TREC 2004 Experiment: Genomics Track

The University of Iowa
Novelty, Question Answering and Genomics: The University of Iowa Response

University of Maryland, Baltimore County
Knowledge-Intensive and Statistical Approaches to the Retrieval and Annotation of Genomics
MEDLINE Citations

University of Maryland, College Park
Knowledge-Intensive and Statistical Approaches to the Retrieval and Annotation of Genomics
MEDLINE Citations

University of Padova
Expanding Queries Using Stems and Symbols

University of Sheffield
Sheffield University and the TREC 2004 Genomics Track:
Query Expansion Using Synonymous Terms

University of Tampere (UTA)
TREC 2004 Genomics Track Experiments at UTA:
The Effects of Primary Keys, Bigram Phrases and Query Expansion on
Retrieval Performance

University of Waterloo
Domain-Specific Synonym Expansion and Validation for Biomedical Information Retrieval
(MultiText Experiments for TREC 2004)

University of Wisconsin, Madison
Exploiting Zone Information, Syntactic Rules, and Informative Terms in Gene Ontology
Annotation of Biomedical Documents

York University
York University at TREC 2004: HARD and Genomics Tracks

HARD

Bilkent University
Approaches to High Accuracy Retrieval:
Phrase-Based Search Experiments in the HARD Track

Chinese Academy of Sciences
ISCAS at TREC 2004: HARD Track

Clairvoyance Corporation
TREC 2004 HARD Track Experiments in Clustering

Indiana University
WIDIT in TREC 2004 Genomics, Hard, Robust and Web Tracks

Johns Hopkins University
Improving Passage Retrieval Using Interactive Elicition and Statistical Modeling

Microsoft Research Ltd
Microsoft Cambridge at TREC 13: Web and Hard Tracks

The Robert Gordon University
The Robert Gordon University's HARD Track Experiments at TREC 2004

Rutgers University
Rutgers' HARD Track Experiences at TREC 2004

University of Chicago
University of Chicago at TREC 2004: HARD Track

University of Illinois at Urbana-Champaign
UIUC in HARD 2004--Passage Retrieval Using HMMs

University of Massachusetts
UMass at TREC 2004: Novelty and HARD

University of Massachusetts, Amherst
HARD Track Overview in TREC 2004
High Accuracy Retrieval from Documents

University of Maryland College Park
Improving Passage Retrieval Using Interactive Elicition and Statistical Modeling

University of North Carolina at Chapel Hill
University of North Carolina's HARD Track Experiments at TREC 2004

Novelty

Question Answering

IBM T.J. Watson Research Center
IBM's PIQUANT II in TREC 2004

ITC-irst
Combining Linguistic Processing and Web Mining for Question Answering: ITC-irst at TREC 2004

Johns Hopkins Applied Physics Lab
Question Answering with QACTIS at TREC 2004

Korea University
Korea University Question Answering System at TREC 2004

LexiClone
LexiClone Inc. and NIST TREC

Macquarie University
AnswerFinder at TREC 2004

MIT Computer Science and Artificial Intelligence Laboratory
Answering Multiple Questions on a Topic From Heterogeneous Resources

National University of Singapore
National University of Singapore at the TREC 13 Question Answering Main Task

National Institute of Standards and Technology
Overview of the TREC 2004 Question Answering Track

Saarland University/DFKI GmbH
Question Answering by Searching Large Corpora With Linguistic Methods

Tsinghua University
THUIR at TREC 2004: QA

University of Amsterdam
Using Wikipedia at the TREC QA Track

University of Edinburgh
Question Answering with QED and Wee at TREC 2004

The University of Iowa
Novelty, Question Answering and Genomics: The University of Iowa Response

University of Lethbridge
University of Lethbridge's Participation in TREC 2004 QA Track

Robust

Terabyte

The Johns Hopkins University Applied Physics Laboratory
JHU/APL at TREC 2004:
Robust and Terabyte Tracks

Microsoft Research
Overview of the TREC 2004 Terabyte Track

Microsoft Research Asia
Microsoft Research Asia at Web Track and Terabyte Track of TREC 2004

National Institute of Standards and Technology (NIST)
Overview of the TREC 2004 Terabyte Track

RMIT University
RMIT University at TREC 2004

Tsinghua University
Using Normal PC to Index and Retrieval Terabyte Document--THUIR at TREC 2004 Terabyte Track

University of Alaska Fairbanks
Towards Grid-Based Information Retrieval

University College Dublin
Experiments in Terabyte Searching, Genomic Retrieval and Novelty Detection for TREC 2004

University of Glasgow
University of Glasgow at TREC 2004:
Experiments in Web, Robust, and Terabyte Tracks with Terrier

University of Massachusetts
Indri at TREC 2004: Terabyte Track

The University of Melbourne
Melbourne University 2004: Terabyte and Web Tracks

Universita di Pisa, Italy
Using Clustering and Blade Clusters in the Terabyte Task

University of Waterloo
Overview of the TREC 2004 Terabyte Track

Web

University of Glasgow
University of Glasgow at TREC 2004:
Experiments in Web, Robust, and Terabyte Tracks with Terrier

The University of Melbourne
Melbourne University 2004: Terabyte and Web Tracks

University of Paris Dauphine
Novel Approaches in Text Information Retrieval
Experiments in the Web Track of TREC 2004

Abstract

This report constitutes the proceedings of the 2004 edition of the Text REtrieval Conference, TREC 2004, held in Gaithersburg, Maryland, November 16–19, 2004. The conference was co-sponsored by the National Institute of Standards and Technology (NIST), the Advanced Research and Development Activity (ARDA), and the Defense Advanced Research Projects Agency (DARPA). TREC 2004 had 103 participating groups including participants from 21 different countries.

TREC 2004 is the latest in a series of workshops designed to foster research in text retrieval and related technologies. This year's conference consisted of seven different tasks: web-based retrieval, novelty detection, question answering, retrieval in the genomics domain, improving the consistency of retrieval systems across queries, improving retrieval effectiveness by focusing on user context, and retrieval from terabyte-scale collections.

The conference included paper sessions and discussion groups. The overview papers for the different "tracks" and for the conference as a whole are gathered in this bound version of the proceedings. The papers from the individual participants and the evaluation output for the runs submitted to TREC 2004 are contained on the disk included in the volume. The TREC 2004 proceedings web site (`http://trec.nist.gov/pubs.html`) also contains the complete proceedings, including system descriptions that detail the timing and storage requirements of the different runs.

Overview of TREC 2004

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899

1 Introduction

The thirteenth Text REtrieval Conference, TREC 2004, was held at the National Institute of Standards and Technology (NIST) November 16–19, 2004. The conference was co-sponsored by NIST, the US Department of Defense Advanced Research and Development Activity (ARDA), and the Defense Advanced Research Projects Agency (DARPA).

TREC 2004 is the latest in a series of workshops designed to foster research on technologies for information retrieval. The workshop series has four goals:

- to encourage retrieval research based on large test collections;
- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

TREC 2004 contained seven areas of focus called "tracks". Six of the tracks had run in at least one previous TREC, while the seventh track, the terabyte track, was new in TREC 2004. The retrieval tasks performed in each of the tracks are summarized in Section 3 below.

Table 2 at the end of this paper lists the 103 groups that participated in TREC 2004. The participating groups come from 21 different countries and include academic, commercial, and government institutions.

This paper serves as an introduction to the research described in detail in the remainder of the volume. The next section provides a summary of the retrieval background knowledge that is assumed in the other papers. Section 3 presents a short description of each track—a more complete description of a track can be found in that track's overview paper in the proceedings. The final section looks toward future TREC conferences.

2 Information Retrieval

Information retrieval is concerned with locating information that will satisfy a user's information need. Traditionally, the emphasis has been on text retrieval: providing access to natural language texts where the set of documents to be searched is large and topically diverse. There is increasing interest, however, in finding appropriate information regardless of the medium that happens to contain that information. Thus "document" can be interpreted as any unit of information such as a web page or a MEDLINE record.

The prototypical retrieval task is a researcher doing a literature search in a library. In this environment the retrieval system knows the set of documents to be searched (the library's holdings), but cannot anticipate the particular topic that will be investigated. We call this an *ad hoc* retrieval task, reflecting the arbitrary subject of the search and its short duration. Other examples of ad hoc searches are web surfers using Internet search engines, lawyers performing patent searches or looking for precedences in case law, and analysts searching archived news reports for particular events. A retrieval system's response to an ad hoc search is generally a list of documents ranked by decreasing similarity to the query. Most of the retrieval tasks in TREC 2004 are ad hoc tasks.

A *known-item search* is similar to an ad hoc search but the target of the search is a particular document (or a small set of documents) that the searcher knows to exist in the collection and wants to find again. Once again, the retrieval system's response is usually a ranked list of documents, and the system is evaluated by the rank at which the target document is retrieved. The named page finding part of the web track task is a known-item search.

In a *categorization* task, the system is responsible for assigning a document to one or more categories from among a given set of categories. The genomics track had several categorization tasks in TREC 2004, and the novelty track tasks required assigning sentences from within documents to "relevant" and "novel" categories. The web track also had a variant of a categorization task, though in this case the topics, not the documents, were to be categorized.

Information retrieval has traditionally focused on returning entire documents that contain answers to questions rather than returning the answers themselves. This emphasis is both a reflection of retrieval systems' heritage as library reference systems and an acknowledgement of the difficulty of question answering. However, for certain types of questions, users would much prefer the system to answer the question than be forced to wade through a list of documents looking for the specific answer. To encourage research on systems that return answers instead of document lists, TREC has had a question answering track since 1999.

2.1 Test collections

Text retrieval has a long history of using retrieval experiments on test collections to advance the state of the art [3, 6, 9], and TREC continues this tradition. A test collection is an abstraction of an operational retrieval environment that provides a means for researchers to explore the relative benefits of different retrieval strategies in a laboratory setting. Test collections consist of three parts: a set of documents, a set of information needs (called *topics* in TREC), and *relevance judgments*, an indication of which documents should be retrieved in response to which topics.

2.1.1 Documents

The document set of a test collection should be a sample of the kinds of texts that will be encountered in the operational setting of interest. It is important that the document set reflect the diversity of subject matter, word choice, literary styles, document formats, etc. of the operational setting for the retrieval results to be representative of the performance in the real task. Frequently, this means the document set must be large. The primary TREC test collections contain about 2 gigabytes of text (between 500,000 and 1,000,000 documents). The document sets used in various tracks have been smaller and larger depending on the needs of the track and the availability of data. The terabyte track was introduced this year to investigate both retrieval and evaluation issues associated with collections significantly larger than 2 gigabytes of text.

The primary TREC document sets consist mostly of newspaper or newswire articles, though there are also some government documents (the *Federal Register*, patent applications) and computer science abstracts (*Computer Selects* by Ziff-Davis publishing) included. High-level structures within each document are tagged using SGML, and each document is assigned an unique identifier called the DOCNO. In keeping of the spirit of realism, the text was kept as close to the original as possible. No attempt was made to correct spelling errors, sentence fragments, strange formatting around tables, or similar faults.

2.1.2 Topics

TREC distinguishes between a statement of information need (the topic) and the data structure that is actually given to a retrieval system (the query). The TREC test collections provide topics to allow a wide range of query construction methods to be tested and also to include a clear statement of what criteria make a document relevant. The format of a topic statement has evolved since the earliest TRECs, but it has been stable since TREC-5 (1996). A topic statement generally consists of four sections: an identifier, a title, a description, and a narrative. An example topic taken from this year's robust track is shown in figure 1.

The different parts of the TREC topics allow researchers to investigate the effect of different query lengths on retrieval performance. For topics 301 and later, the "title" field was specially designed to allow experiments with very

```
<num> Number:  656
<title> lead poisoning children
<desc>
How are young children being protected against lead poisoning from paint and
water pipes?
<narr>
Documents describing the extent of the problem, including suits against
manufacturers and product recalls, are relevant.  Descriptions of future plans
for lead poisoning abatement projects are also relevant.  Worker problems with
lead are not relevant.  Other poison hazards for children are not relevant.
```

Figure 1: A sample TREC 2004 topic from the robust track test set.

short queries; these title fields consist of up to three words that best describe the topic. The description ("desc") field is a one sentence description of the topic area. The narrative ("narr") gives a concise description of what makes a document relevant.

Participants are free to use any method they wish to create queries from the topic statements. TREC distinguishes among two major categories of query construction techniques, automatic methods and manual methods. An automatic method is a means of deriving a query from the topic statement with no manual intervention whatsoever; a manual method is anything else. The definition of manual query construction methods is very broad, ranging from simple tweaks to an automatically derived query, through manual construction of an initial query, to multiple query reformulations based on the document sets retrieved. Since these methods require radically different amounts of (human) effort, care must be taken when comparing manual results to ensure that the runs are truly comparable.

TREC topic statements are created by the same person who performs the relevance assessments for that topic (the *assessor*). Usually, each assessor comes to NIST with ideas for topics based on his or her own interests, and searches the document collection using NIST's PRISE system to estimate the likely number of relevant documents per candidate topic. The NIST TREC team selects the final set of topics from among these candidate topics based on the estimated number of relevant documents and balancing the load across assessors.

2.1.3 Relevance judgments

The relevance judgments are what turns a set of documents and topics into a test collection. Given a set of relevance judgments, the retrieval task is then to retrieve all of the relevant documents and none of the irrelevant documents. TREC usually uses binary relevance judgments—either a document is relevant to the topic or it is not. To define relevance for the assessors, the assessors are told to assume that they are writing a report on the subject of the topic statement. If they would use any information contained in the document in the report, then the (entire) document should be marked relevant, otherwise it should be marked irrelevant. The assessors are instructed to judge a document as relevant regardless of the number of other documents that contain the same information.

Relevance is inherently subjective. Relevance judgments are known to differ across judges and for the same judge at different times [7]. Furthermore, a set of static, binary relevance judgments makes no provision for the fact that a real user's perception of relevance changes as he or she interacts with the retrieved documents. Despite the idiosyncratic nature of relevance, test collections are useful abstractions because the *comparative* effectiveness of different retrieval methods is stable in the face of changes to the relevance judgments [10].

The relevance judgments in early retrieval test collections were complete. That is, a relevance decision was made for every document in the collection for every topic. The size of the TREC document sets makes complete judgments utterly infeasible—with 800,000 documents, it would take over 6500 hours to judge the entire document set for one topic, assuming each document could be judged in just 30 seconds. Instead, TREC uses a technique called pooling [8] to create a subset of the documents (the "pool") to judge for a topic. Each document in the pool for a topic is judged for relevance by the topic author. Documents that are not in the pool are assumed to be irrelevant to that topic.

The judgment pools are created as follows. When participants submit their retrieval runs to NIST, they rank their runs in the order they prefer them to be judged. NIST chooses a number of runs to be merged into the pools, and selects that many runs from each participant respecting the preferred ordering. For each selected run, the top X documents (usually, $X = 100$) per topic are added to the topics' pools. Since the retrieval results are ranked by decreasing similarity to the query, the top documents are the documents most likely to be relevant to the topic. Many documents are retrieved in the top X for more than one run, so the pools are generally much smaller than the theoretical maximum of X × *the-number-of-selected-runs* documents (usually about 1/3 the maximum size).

The use of pooling to produce a test collection has been questioned because unjudged documents are assumed to be not relevant. Critics argue that evaluation scores for methods that did not contribute to the pools will be deflated relative to methods that did contribute because the non-contributors will have highly ranked unjudged documents.

Zobel demonstrated that the quality of the pools (the number and diversity of runs contributing to the pools and the depth to which those runs are judged) does affect the quality of the final collection [14]. He also found that the TREC collections were not biased against unjudged runs. In this test, he evaluated each run that contributed to the pools using both the official set of relevant documents published for that collection and the set of relevant documents produced by removing the relevant documents uniquely retrieved by the run being evaluated. For the TREC-5 ad hoc collection, he found that using the unique relevant documents increased a run's 11 point average precision score by an average of 0.5 %. The maximum increase for any run was 3.5 %. The average increase for the TREC-3 ad hoc collection was somewhat higher at 2.2 %.

A similar investigation of the TREC-8 ad hoc collection showed that every automatic run that had a mean average precision score of at least 0.1 had a percentage difference of less than 1 % between the scores with and without that group's uniquely retrieved relevant documents [13]. That investigation also showed that the quality of the pools is significantly enhanced by the presence of recall-oriented manual runs, an effect noted by the organizers of the NTCIR (NACSIS Test Collection for evaluation of Information Retrieval systems) workshop who performed their own manual runs to supplement their pools [5].

While the lack of any appreciable difference in the scores of submitted runs is not a guarantee that all relevant documents have been found, it is very strong evidence that the test collection is reliable for comparative evaluations of retrieval runs. The differences in scores resulting from incomplete pools observed here are smaller than the differences that result from using different relevance assessors [10].

2.2 Evaluation

Retrieval runs on a test collection can be evaluated in a number of ways. In TREC, ad hoc tasks are evaluated using the `trec_eval` package written by Chris Buckley of Sabir Research [1]. This package reports about 85 different numbers for a run, including *recall* and *precision* at various cut-off levels plus single-valued summary measures that are derived from recall and precision. Precision is the proportion of retrieved documents that are relevant, while recall is the proportion of relevant documents that are retrieved. A cut-off level is a rank that defines the retrieved set; for example, a cut-off level of ten defines the retrieved set as the top ten documents in the ranked list. The `trec_eval` program reports the scores as averages over the set of topics where each topic is equally weighted. (The alternative is to weight each relevant document equally and thus give more weight to topics with more relevant documents. Evaluation of retrieval effectiveness historically weights topics equally since all users are assumed to be equally important.)

Precision reaches its maximal value of 1.0 when only relevant documents are retrieved, and recall reaches its maximal value (also 1.0) when all the relevant documents are retrieved. Note, however, that these theoretical maximum values are not obtainable as an average over a set of topics at a single cut-off level because different topics have different numbers of relevant documents. For example, a topic that has fewer than ten relevant documents will have a precision score less than one at ten documents retrieved regardless of how the documents are ranked. Similarly, a topic with more than ten relevant documents must have a recall score less than one at ten documents retrieved. At a single cut-off level, recall and precision reflect the same information, namely the number of relevant documents retrieved. At varying cut-off levels, recall and precision tend to be inversely related since retrieving more documents will usually increase recall while degrading precision and vice versa.

Of all the numbers reported by `trec_eval`, the recall-precision curve and mean (non-interpolated) average precision are the most commonly used measures to describe TREC retrieval results. A recall-precision curve plots precision as a function of recall. Since the actual recall values obtained for a topic depend on the number of relevant documents, the average recall-precision curve for a set of topics must be interpolated to a set of standard recall values. The particular interpolation method used is given in Appendix A, which also defines many of the other evaluation measures reported by `trec_eval`. Recall-precision graphs show the behavior of a retrieval run over the entire recall spectrum.

Mean average precision is the single-valued summary measure used when an entire graph is too cumbersome. The average precision for a single topic is the mean of the precision obtained after each relevant document is retrieved (using zero as the precision for relevant documents that are not retrieved). The mean average precision for a run consisting of multiple topics is the mean of the average precision scores of each of the individual topics in the run. The average precision measure has a recall component in that it reflects the performance of a retrieval run across all relevant documents, and a precision component in that it weights documents retrieved earlier more heavily than documents retrieved later. Geometrically, average precision is the area underneath a non-interpolated recall-precision curve.

As TREC has expanded into tasks other than the traditional ad hoc retrieval task, new evaluation measures have had to be devised. Indeed, developing an appropriate evaluation methodology for a new task is one of the primary goals of the TREC tracks. The details of the evaluation methodology used in a track are described in the track's overview paper.

3 TREC 2004 Tracks

TREC's track structure was begun in TREC-3 (1994). The tracks serve several purposes. First, tracks act as incubators for new research areas: the first running of a track often defines what the problem *really* is, and a track creates the necessary infrastructure (test collections, evaluation methodology, etc.) to support research on its task. The tracks also demonstrate the robustness of core retrieval technology in that the same techniques are frequently appropriate for a variety of tasks. Finally, the tracks make TREC attractive to a broader community by providing tasks that match the research interests of more groups.

Table 1 lists the different tracks that were in each TREC, the number of groups that submitted runs to that track, and the total number of groups that participated in each TREC. The tasks within the tracks offered for a given TREC have diverged as TREC has progressed. This has helped fuel the growth in the number of participants, but has also created a smaller common base of experience among participants since each participant tends to submit runs to fewer tracks.

This section describes the tasks performed in the TREC 2004 tracks. See the track reports later in these proceedings for a more complete description of each track.

3.1 The genomics track

The genomics track was introduced as a "pre-track" in 2002. It is the first TREC track devoted to retrieval within a specific domain; one of the goals of the track is to see how exploiting domain-specific information improves retrieval effectiveness.

The 2004 genomics track contained an ad hoc retrieval task and three variants of a categorization task. The ad hoc task used a 10-year subset (1994–2003) of MEDLINE, a bibliographic database of the biomedical literature maintained by the US National Library of Medicine who donated the subset to the track. The subset used in the track contains about 4.5 million MEDLINE records (which include title and abstract as well as other bibliographic information) and is about 9GB of data. The 50 topics for the ad hoc task were derived from information needs obtained through interviews of biomedical researchers. Pools were created using one run from each of the 27 participating groups using a depth of 75. Relevance judgments were made by assessors with backgrounds in biology using a three-point scale of definitely relevant, probably relevant, and not relevant. Both definitely relevant and probably relevant were considered relevant when computing evaluation scores.

Domain knowledge was most frequently exploited by using resources such as the MeSH hierarchy (a controlled vocabulary used to index medical literature) to expand queries. Careful use of such resources appears to increase retrieval effectiveness, though some attempts to exploit such information decreased effectiveness relative to a generic baseline.

The genomics domain has a number of model organism database projects in which the literature regarding a specific organism (such as a mouse) is tracked and annotated with the function of genes and proteins. The classification tasks

Table 1: Number of participants per track and total number of distinct participants in each TREC

Track	TREC												
	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Ad Hoc	18	24	26	23	28	31	42	41	—	—	—	—	—
Routing	16	25	25	15	16	21	—	—	—	—	—	—	—
Interactive	—	—	3	11	2	9	8	7	6	6	6	—	—
Spanish	—	—	4	10	7	—	—	—	—	—	—	—	—
Confusion	—	—	—	4	5	—	—	—	—	—	—	—	—
DB Merging	—	—	—	3	3	—	—	—	—	—	—	—	—
Filtering	—	—	—	4	7	10	12	14	15	19	21	—	—
Chinese	—	—	—	—	9	12	—	—	—	—	—	—	—
NLP	—	—	—	—	4	2	—	—	—	—	—	—	—
Speech	—	—	—	—	—	13	10	10	3	—	—	—	—
Cross-Language	—	—	—	—	—	13	9	13	16	10	9	—	—
High Precision	—	—	—	—	—	5	4	—	—	—	—	—	—
VLC	—	—	—	—	—	—	7	6	—	—	—	—	—
Query	—	—	—	—	—	—	2	5	6	—	—	—	—
QA	—	—	—	—	—	—	—	20	28	36	34	33	28
Web	—	—	—	—	—	—	—	17	23	30	23	27	18
Video	—	—	—	—	—	—	—	—	—	12	19	—	—
Novelty	—	—	—	—	—	—	—	—	—	—	13	14	14
Genomics	—	—	—	—	—	—	—	—	—	—	—	29	33
HARD	—	—	—	—	—	—	—	—	—	—	—	14	16
Robust	—	—	—	—	—	—	—	—	—	—	—	16	14
Terabyte	—	—	—	—	—	—	—	—	—	—	—	—	17
Total participants	22	31	33	36	38	51	56	66	69	87	93	93	103

used in the 2004 track mimic some aspects of this curation process with the goal of eventually automating this now largely manual task. For the classification tasks, the track used the full text articles from a two-year span of three journals. This text was made available to the track through Highwire Press. The truth data for the tasks came from the actual annotation process carried out by the human annotators in the mouse genome informatics (MGI) system. Evaluation scores were computed using normalized utility measures.

As in the ad hoc task, many groups used MeSH terms as features to classify the documents. While these approaches were relatively effective, a subsequent analysis demonstrated the benefit was largely attributable to a single MeSH term: a baseline run that classified documents solely by the presence of the MeSH term *Mice* in the MEDLINE record of the document would have been the second best run submitted to the track for the triage classification task.

3.2 The HARD track

HARD stands for "High Accuracy Retrieval from Documents". The HARD track was started in TREC 2003 with the goal of improving retrieval performance, especially at the top of the ranked list, by targeting retrieval results to the specific searcher. To facilitate such targeting, the HARD track provides metadata in the topic statement. In addition, "clarification forms" provide a limited means of interaction between the system and the searcher.

The underlying task in the HARD track was an ad hoc retrieval task. The document set was a set of newswire/newspaper articles from 2003, including (English portions) of non-US papers. The collection is approximately 1500MB of text and contains approximately 650,000 articles. Topics were created at the Linguistic Data Consortium (LDC), and were originally released in standard TREC format (i.e., just title, description, and narrative fields). Once participants submitted baseline runs using the standard topics, they received the expanded version of the topics. There were 50 topics in the test set, though only 45 topics were used in the evaluation since five topics had no relevant documents.

The expanded version of the topics contained both a statement of the retrieval unit and the metadata. The retrieval unit was always specified, and was either "passage" or "document". The "passage" specification meant retrieval systems should return pieces of documents, rather than full documents, as a response. The types of metadata in the TREC 2004 topics included familiarity, genre, geography, subject, and related text. The first three types affected the relevance of a text: a text that was on-topic but did not satisfy one of these metadata constraints was considered not relevant when using stringent relevance criteria. The subject metadata item contained the subject domain of the topic (for example, "sports", or "politics"); a document that did not meet this criterion was off-topic. The related text metadata provided some examples of relevant or on-topic text drawn from outside the test corpus. Different topics contained different kinds and amounts of metadata.

In addition to the information included in the expanded version of the topics, participants could collect information from the searcher (the assessor who created and judged the topic) using clarification forms. A clarification form was a single, self-contained HTML form created by the participating group and specific to a single topic. There were no restrictions on what type of data could be collected using a clarification form, but the searcher spent no more than three minutes filling out any one form.

Participants then made new runs using any combination of information from the expanded topics and clarification forms. The goal was to see if the additional information helped systems to create a more effective retrieved set than the initial baseline result. Retrieval results were evaluated both at the document level (for all 45 topics including those with retrieval unit "passage") using `trec_eval` and using passage level evaluation measures over just the 25 topics with retrieval unit "passage".

Sixteen groups submitted 135 runs to the HARD track. Most groups were able to exploit the additional information to improve effectiveness as compared to their baseline run, generally by performing some type of relevance feedback.

3.3 The novelty track

The goal of the novelty track is to investigate systems' abilities to locate relevant and new (nonredundant) information within an ordered set of documents. This task models an application where the user is skimming a set of documents and the system highlights the new, on-topic information. The track was first introduced in TREC 2002, though the tasks changed significantly between 2002 and 2003. This year's track used the same tasks as the 2003 track.

The basic task in the novelty track is as follows: given a topic and an ordered set of documents segmented into sentences, return sentences that are both relevant to the topic and novel given what has already been seen. To accomplish this task, participants must first identify relevant sentences and then identify which sentences contain new information.

Fifty new topics were created for the 2004 track. As in TREC 2003, half of the topics focused on events and the other half focused on opinions about controversial subjects. For each topic, the assessor created a statement of information need and queried the document collection using the NIST PRISE search engine. The assessor selected 25 relevant documents and labeled the relevant and new sentences in each. The document collection used was the *AQUAINT Corpus of English News Text* which contains approximately 1,033,000 documents and 3 gigabytes of text. The document set for a topic in the test set contained the 25 relevant documents selected by the assessor as well as 0 or more irrelevant documents. The documents in a set were ordered chronologically.

There were four tasks in the track, which allowed participants to test their approaches to novelty detection using no, partial, or complete relevance information.

Task 1. Given the complete document set for a topic, identify all relevant and novel sentences.

Task 2. Given the relevant sentences in the complete document set, identify all novel sentences.

Task 3. Given the relevant and novel sentences in the first 5 documents for the topic, find the relevant and novel sentences in the remaining documents.

Task 4. Given the relevant sentences in the complete document set, and the novel sentences in the first 5 documents, find the novel sentences in the remaining documents.

Given the set of relevant and new sentences selected by the assessor who created the topic, the score for a novelty topic was computed as the F measure where sentence set recall and sentence set precision are equally weighted.

Fourteen groups submitted 183 runs to the novelty track, with tasks 1 and 2 having the greater participation. The inclusion of nonrelevant documents in the retrieved set appears to make task 1 much more challenging. In TREC 2003,

3	Hale Bopp comet	
3.1	FACTOID	When was the comet discovered?
3.2	FACTOID	How often does it approach the earth?
3.3	LIST	In what countries was the comet visible on its last return?
3.4	OTHER	

Figure 2: A sample QA track question series.

the best-performing systems for task 1 were roughly comparable to human performance as measured by scoring a second assessor's sentence selection against the primary assessor's choices. This year, the best systems' effectiveness was well below human performance. The particular topics used this year may also have been more difficult given that the absolute scores of TREC 2004 systems were lower than TREC 2003 scores for task 2 and task 2 is unaffected by nonrelevant documents.

3.4 The question answering (QA) track

The question answering track addresses the problem of information overload by encouraging research into systems that return actual answers, as opposed to ranked lists of documents, in response to a question. The TREC 2003 version of the track used a combined task where the test set of questions consisted of factoid, list, and definition questions. Each type of question was judged and scored separately, but the final score for a run was a weighted average of the component scores. The task in the 2004 track was similar in that the test set consisted of a mix of question types, and the final score was a weighted average of the components. The task was reorganized, however, such that the systems were to answer a series of factoid and list questions that each related to a common target, and then to respond with a list of "other" information about the target that was not covered by the previous questions in the series. This last question in the series is a more difficult variant of the definition questions in TREC 2003. This reorientation of the task requires systems to track context when answering questions, an important element of question answering that the track has not yet successfully incorporated [11].

The document set used in the track was the *AQUAINT Corpus of English News Text*. The test set consisted of 65 series of questions that together included 230 factoid questions, 56 list questions (one had to be removed from the evaluation due to no correct answers in the collection), and 65 Other questions (one had to be removed from the evaluation since it mistakenly went unjudged). Each of the questions was explicitly tagged as to what type of question it was and what series it belonged to. The target of the series was given as metadata for the whole series. An example series is given in figure 2.

The score for the factoid question component was accuracy, the percentage of factoid questions whose response was judged correct. The list and Other question components were each scored using average F, though the computation of the F score differed between the two components [12]. The final score for a run was computed as a weighted average of the three component scores: $\text{FinalScore} = .5\text{Accuracy} + .25\text{AveListF} + .25\text{AveOtherF}$.

Sixty-three runs from 28 different groups were submitted to the track. In general, the use of pronouns and anaphora in questions later in a series did not seem to pose a very serious challenge for the systems, in part because the target was the correct referent a large majority of the time. For most systems, the average score for the first question in a series was somewhat greater than the average score for a question that was not the first question in a series, but the difference was not great and is confounded by other effects (there are many fewer first questions to compute the average over, first questions in a series might be intrinsically easier questions, etc.).

The reorganization of the task into a set of question series had an unexpected benefit. The series proved to be an appropriate level of granularity for aggregating scores for an effective evaluation. The series is small enough to be meaningful at the task level since it represents a single user interaction, yet it is large enough to avoid the highly skewed score distributions exhibited by single questions. Computing a combined score for each series, and averaging the series scores, produces a QA task evaluation that more closely mimics classic document retrieval evaluation.

3.5 The robust track

The robust track looks to improve the consistency of retrieval technology by focusing on poorly performing topics. TREC 2004 was the second time the track was run. The initial track provided strong evidence that optimizing average effectiveness using the standard methodology and current evaluation measures further improves the effectiveness of the already-effective topics, sometimes at the expense of the poor performers. That track also showed that measuring poor performance is intrinsically difficult because there is so little signal in the sea of noise for a poorly performing topic. New measures devised for the TREC 2003 robust track do emphasize poorly performing topics, but because there is so little information, the measures are unstable.

The task in both years of the robust track was a classic ad hoc retrieval task. The TREC 2004 edition of the track used more topics than the 2003 edition in hopes of getting a more stable evaluation. In particular, the test set for 2004 consisted of 250 topics (one topic was dropped from the evaluation since it was judged to have no relevant documents). Two hundred of the topics were used in previous TREC tasks and 50 new topics were created for the track. To avoid needing new relevance judgments for the 200 old topics, an old document set was used: the set of documents on TREC disks 4 and 5 minus the *Congressional Record* documents.

The use of old topics had an additional motivation other than not needing new relevance judgments for those topics. Since the retrieval results from the previous TREC in which the topics were used are available, it is possible to select topics that are known to be challenging to a majority of retrieval systems. Fifty topics from among the 200 old topics were designated as being difficult. These topics were selected for the TREC 2003 track by choosing topics that had a low median average precision score and at least one high outlying score.

The retrieval results were evaluated using `trec_eval`, two measures introduced in the TREC 2003 track that emphasize poorly performing topics, and a new measure, geometric MAP, introduced in this year's track. The geometric MAP is a variant of the traditional MAP measure that uses a geometric mean rather than an arithmetic mean to average individual topic results. An analysis of the behavior of the geometric MAP measure suggests it gives appropriate emphasis to poorly performing topics while being more stable at equal topic set sizes.

The robust track received a total of 110 runs from 14 participants. All of the runs submitted to the track were automatic runs. The results indicate that the most promising approach to improving poorly performing topics is exploiting text collections other than the target collection, though the process must be carefully controlled to avoid making the results worse. The web was the collection most frequently used as an auxiliary collection.

An additional requirement in this year's track was for systems to submit a ranked list of the topics ordered by perceived difficulty. That is, the system assigned each topic a number from 1 to 250 where the topic assigned 1 was the topic the system believed it did best on, the topic assigned 2 was the topic the system believed it did next best on, etc. The purpose of the requirement was to see if systems can recognize whether a topic is difficult at run time, a first step toward doing special processing for difficult topics. While some systems were clearly better than others at predicting when a topic is difficult for that system, none of the systems were particularly good at the task. How much accuracy is required to make effective use of the predictions is still unknown.

3.6 The terabyte track

The terabyte track is a new track in 2004. The goal of the track is is to develop an evaluation methodology for terabyte-scale document collections. The track also provides an opportunity for participants to see how well their retrieval algorithms scale to much larger test sets than other TREC collections.

The document collection used in the track is the GOV2 collection, a collection of Web data crawled from Web sites in the .gov domain during early 2004. This collection contains a large proportion of the crawlable pages in .gov, including html and text, plus extracted text of pdf, word and postscript files. The collection is 426GB in size and contains approximately 25 million documents. The collection is smaller than a full terabyte due to the difficulty of obtaining and processing enough documents while allowing sufficient time for distributing the collection to participants. The collection will be expanded using data from other sources in future years. The current collection is at least an order of magnitude greater than the next-largest TREC collection.

The task in the track was a classic ad hoc retrieval task. The test set consisted of 50 topics created specifically for the track. While the document set consists of web pages, the topics were standard information-seeking requests, and not navigational requests or topic distillation requests, for example. Systems returned the top 10,000 documents per topic so various evaluation strategies can be investigated. Participants also answered a series of questions about timing and resources required to produce the retrieval results.

Seventy runs from 17 different groups were submitted to the track. The top 85 documents per topic for two runs per group were added to the judgment pools. Initial analysis of the track results has revealed little difference in the relative effectiveness of different approaches when evaluated by MAP or by bpref, a measure created for evaluation environments where pools are known to be very incomplete [2]. There are a variety of reasons why this might be so: it may mean that current pooling practices are adequate for collections of this size, or that the runs submitted to the terabyte track happened to retrieve a sufficient set of relevant documents, or that the terabyte topics happened to be particularly narrow, and so forth. The terabyte track will continue in TREC 2005 to examine these questions.

3.7 The web track

The goal in the web track is to investigate retrieval behavior when the collection to be searched is a large hyperlinked structure such as the World Wide Web. Previous TREC web tracks had separately investigated topic distillation, named page finding, and home page finding tasks [4]. Since web search engines must process these types of searches (among others) without explicit knowledge of which type of search is wanted, this year's web task combined them into a single task.

For a topic distillation search a system is to return a list of entry points for good websites principally devoted to the topic. Since there are only a few good websites for any particular topic, there are only a few key ("relevant") pages for a topic distillation search. The emphasis is on returning entry pages rather than pages containing relevant information themselves since a result list of homepages provides a better overview of the coverage of a topic in the collection.

Named page and home page finding searches are similar to each other in that both are known-item tasks where the system is to return a particular page. For home page finding, the target page is the home page of the entity in the topic. For named page finding, a particular page is sought, but that page is not an entry point to a site (e.g., "1040 tax form").

For the TREC 2004 task, participants received a set of 225 title-only topics such as "West Indian manatee information" and "York county". The assessor specified which type of search was intended when the topic was created, but the test set did not include this information. Systems returned a ranked list of up to 1000 pages per topic. During judging, the assessors made binary judgments as to whether a page was appropriate with respect to the intended task. That is, the pages returned for topics whose search type was topic distillation were judged relevant if the page was a key entry page and not relevant otherwise. For the named page finding and home page finding topics, a page was judged relevant if and only if the page was the target page (or a mirror/alias of the target page). The runs were evaluated using MAP, which is equivalent to the mean reciprocal rank (MRR) measure for known-item searches.

The track used the .GOV collection created for the TREC 2002 web track and distributed by CSIRO. This collection is based on a January, 2002 crawl of .gov web sites. The documents in the collection contain both page content and the information returned by the http daemon; text extracted from the non-html pages is also included in the collection.

In addition to the search task, the track also contained a classification task in which the goal was simply to label each of the 225 test topics as to what type of search was intended.

Eighteen groups submitted a total of 83 runs to the track. Nine of the runs were classification task runs. The retrieval results showed that systems are able to obtain effective overall retrieval without having to classify the queries by type. That is, groups were able to devise a single technique that performed well for home page, named page, and distillation topics. These techniques were not based solely on the text of a page, but also needed to exploit some sort of web information such as link structure or anchor text. Systems that did attempt to classify topics were generally able to do so, with most classification errors confusing named page and home page topics.

4 The Future

A significant fraction of the time of one TREC workshop is spent in planning the next TREC. A majority of the TREC 2004 tracks will continue in TREC 2005, including the genomics, HARD, QA, robust, and terabyte tracks. As described in the web track overview paper, the web track as such will end, with a new enterprise track taking its place. The goal of the enterprise track is to study enterprise search—satisfying a user who is searching the data of an organization to accomplish some task. The novelty track will also end. Finally, a new track, the spam track, will be introduced in TREC 2005. The goal of the spam track is to provide a standard evaluation of current and proposed spam filtering approaches, thereby laying the foundation for the evaluation of more general email filtering and retrieval tasks.

Acknowledgements

Special thanks to the track coordinators who make the variety of different tasks addressed in TREC possible.

References

[1] Chris Buckley. trec_eval IR evaluation package. Available from `http://trec.nist.gov/trec_eval/`.

[2] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, 2004.

[3] C. W. Cleverdon, J. Mills, and E. M. Keen. Factors determining the performance of indexing systems. Two volumes, Cranfield, England, 1968.

[4] Nick Craswell, David Hawking, Ross Wilkinson, and Mingfang Wu. Overview of the TREC 2003 web track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 78–92, 2004.

[5] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. Overview of IR tasks at the first NTCIR workshop. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 11–44, 1999.

[6] G. Salton, editor. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1971.

[7] Linda Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology*, 29:3–48, 1994.

[8] K. Sparck Jones and C. van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

[9] Karen Sparck Jones. *Information Retrieval Experiment*. Butterworths, London, 1981.

[10] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716, 2000.

[11] Ellen M. Voorhees. Overview of the TREC 2001 question answering track. In *Proceedings of the Tenth Text REtreival Conference (TREC 2001)*, pages 42–51, 2002.

[12] Ellen M. Voorhees. Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 54–68, 2004.

[13] Ellen M. Voorhees and Donna Harman. Overview of the eighth Text REtrieval Conference (TREC-8). In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 1–24, 2000. NIST Special Publication 500-246. Electronic version available at `http://trec.nist.gov/pubs.html`.

[14] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998. ACM Press, New York.

Table 2: Organizations participating in TREC 2004

Alias-i, Inc.	Arizona State University
California State U. San Marcos	Carnegie Mellon University
Chinese Academy of Sciences (3 groups)	Chinese University of Hong Kong
Clairvoyance Corporation	CL Research
Columbia University	ConverSpeech LLC & Stanford SGD
CSIRO	Dalhousie University
Decision Aid team-LAMSADE	Dublin City University
Etymon	Fondazione Ugo Bordoni
Fudan University (2 groups)	German University in Cairo
Hong Kong Polytechnic University	Hummingbird
IBM India Research Lab	IBM Research Lab Haifa
IBM T.J. Watson Research Center	IDA/CCS/NSA
IIT Information Retrieval Lab	Indiana University (2 groups)
IRIT/SIG	ITC-irst
Johns Hopkins University	Korea University
Language Computer Corporation	LexiClone
Macquarie University	Massachusetts Institute of Technology
Max-Planck-Institute for Computer Science	Meiji University
Microsoft Research Asia	Microsoft Research Ltd
Monash University	National Central University
National Security Agency	National Taiwan University
National University of Singapore	National U. of Singapore & Singapore-MIT Alliance
NLM-UMaryland Team	Oregon Health and Science University
PATOLIS Corporation	Peking University
Queens College, CUNY	RMIT University
Rutgers University (2 groups)	Saarland University
Sabir Research, Inc.	Shanghai JiaoTong University
SUNY at Buffalo	Tarragon Consulting Corporation
The MITRE Corporation	The Robert Gordon University
The University of Melbourne	TNO & Erasmus MC
Tsinghua University (2 groups)	UC Berkeley
U. Hospital Geneva & Swiss Federal Inst. of Tech.	Universidade de Lisboa Campo Grande
Universitat Politcnica de Catalunya	Universit Paris Sud
University of Alaska Fairbanks	University of Alberta
University of Amsterdam	University of Chicago
University of Cincinnati	University of Edinburgh
University of Edinburgh & Sydney	University of Glasgow
University of Illinois at Chicago	University of Illinois at Urbana-Champaign
University of Iowa	University of Lethbridge
University of Limerick	University of Maryland UMIACS
University of Massachusetts	University of Michigan
University of North Carolina	University of North Texas
University of Padova	University of Pisa
University of Sheffield	University of Tampere
University of Tokyo	University of Twente
University of Wales, Bangor	University of Waterloo (2 groups)
University of Wisconsin	USC-Information Sciences Institute
Virginia Tech	York University

TREC 2004 Genomics Track Overview

William R. Hersh[1], Ravi Teja Bhuptiraju[1], Laura Ross[1], Phoebe Johnson[2], Aaron M. Cohen[1], Dale F. Kraemer[1]
[1]Oregon Health & Science University, Portland, OR, USA
[2]Biogen Idec Corp., Cambridge, MA

The TREC 2004 Genomics Track consisted of two tasks. The first task was a standard ad hoc retrieval task using topics obtained from real biomedical research scientists and documents from a large subset of the MEDLINE bibliographic database. The second task focused on categorization of full-text documents, simulating the task of curators of the Mouse Genome Informatics (MGI) system and consisting of three subtasks. One subtask focused on the triage of articles likely to have experimental evidence warranting the assignment of GO terms, while the other two subtasks focused on the assignment of the three top-level GO categories. The track had 33 participating groups.

1. Motivations and Background

The goal of the TREC Genomics Track is to create test collections for evaluation of information retrieval (IR) and related tasks in the genomics domain. The Genomics Track differs from all other TREC tracks in that it is focused on retrieval in a specific domain as opposed to general retrieval tasks, such as Web searching or question answering.

To date, the track has focused on advanced users accessing the scientific literature. The advanced users include biomedical scientists and database curators or annotators. New advances in biotechnologies have changed the face of biological research, particularly "high-throughput" techniques such as gene microarrays [1]. These not only generate massive amounts of data but also have led to an explosion of new scientific knowledge. As a result, this domain is ripe for improved information access and management.

The scientific literature plays a key role in the growth of biomedical research data and knowledge. Experiments identify new genes, diseases, and other biological processes that require further investigation. Furthermore, the literature itself becomes a source of "experiments" as researchers turn to it to search for knowledge that drives new hypotheses and research. Thus there are considerable challenges not only for better IR systems, but also for improvements in related techniques, such as information extraction and text mining [2].

Because of the growing size and complexity of the biomedical literature, there is increasing effort devoted to structuring knowledge in databases. The use of these databases is made pervasive by the growth of the Internet and Web as well as a commitment of the research community to put as much data as possible into the public domain. Figure 1 depicts the overall process of "funneling" the literature to structure knowledge, showing the information system tasks used at different levels along the way. This figure shows our view of the optimal uses for IR and the related areas of information extraction and text mining.

One of the many key efforts is to annotate the function of genes. To facilitate this, the research community has come together to develop the Gene Ontology (GO, www.geneontology.org) [3]. While the GO is not an ontology in the purists' sense, it is a large, controlled vocabulary based on three axes or hierarchies:

- Molecular function - the activity of the gene product at the molecular (biochemical) level, e.g. protein binding
- Biological process - the biological activity carried out by the gene process, e.g., cell differentiation
- Cellular component - where in the cell the gene product functions, e.g., the nucleus

A major use of the GO has been to annotate the genomes of organisms used in biological research. The annotations are often linked to other information, such as literature, the gene sequence, the structure of the resulting protein, etc.. An increasingly common approach is to develop "model organism databases" that bring together all this information in an easy to use format. Some of the better known model organism databases include those devoted to the mouse (Mouse Genome Informatics, MGI,

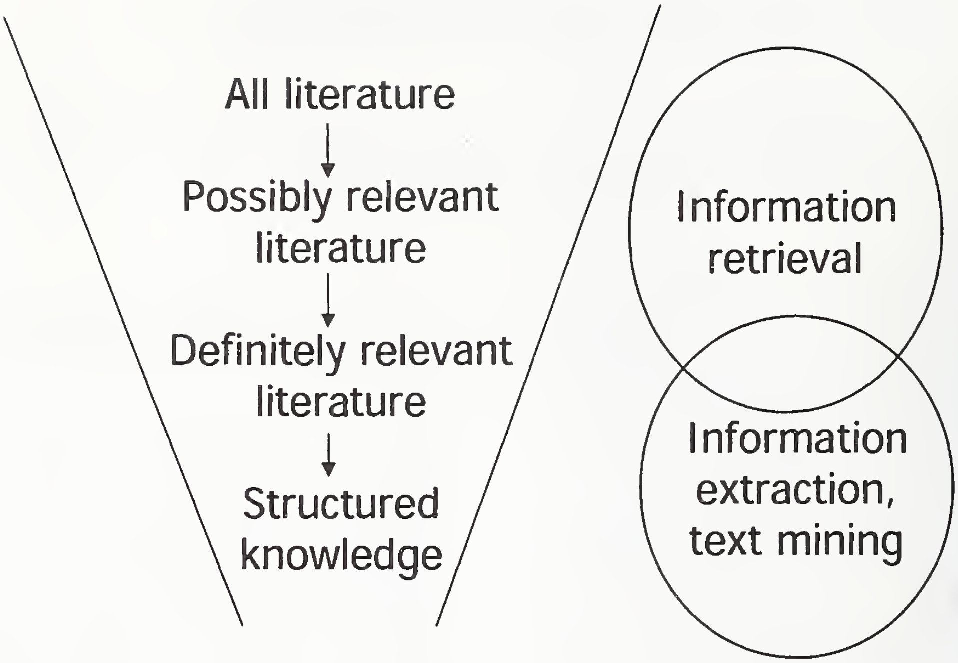

All literature
Possibly relevant literature
Definitely relevant literature
Structured knowledge
Information retrieval
Information extraction, text mining

Figure 1 - The steps in deriving knowledge from the biomedical literature and the associated information systems used along the way.

www.informatics.jax.org) and the yeast (Saccharomyces Genome Database, SGD, www.yeastgenome.org). These databases require extensive human effort for annotation or curation, which is usually done by PhD-level researchers.

These curators could be aided substantially by high-quality information tools, including IR systems.

The 2004 track was the second year of the TREC Genomics Track. This year was different from the first year, as we had resources available to us from a National Science Foundation (NSF) Information Technology Research (ITR) grant that allowed for programming support and relevance judgments. In contrast, for the 2003 track we had to rely on proxies for relevance judgments and other gold standard data [4].

The Genomics Track is overseen by a steering committee of individuals with a background in IR and/or genomics. In early 2003, the committee produced a "road map" that called for modifying one experimental "facet" each year. For the purposes of the roadmap (based on the NSF grant proposal), the original year (2003) was Year 0, making 2004 Year 1. The original plan was to add new types of content in Year 1 and new types of information needs in Year 2. Because we were unable to secure substantial numbers of full text documents for the ad hoc retrieval task in 2004, we decided to reverse the order of the roadmap for Years 1 and 2. This meant we focused on new types of information needs for 2004 (and hopefully new types of content in 2005). However, it should be noted that even in this era of virtually all biomedical journals being available electronically, most users of the literature start their searches using MEDLINE.

2. Overview of Track

In TREC 2004, the Genomics Track had two tasks, the second of which was subdivided into subtasks. The first task was a standard ad hoc retrieval task using topics obtained from surveying real research scientists and searching in a large subset of the MEDLINE bibliographic database. The second task focused on categorization of full-text documents, simulating the task of curators for the MGI system. One subtask focused on the triage of articles likely to have experimental evidence warranting the assignment of GO terms, while the other two subtasks focused on the assignment of the three GO categories (indicating the assignment of a term within them).

A total of 145 runs were submitted for scoring. There were 47 runs from 27 groups submitted for the ad hoc task. There were 98 runs submitted from 20 groups for the categorization task. These were distributed across the subtasks of the categorization task as follows: 59 for the triage subtask, 36 for the annotation hierarchy subtask, and three for the annotation hierarchy plus evidence code subtask. A total of 33 groups participated in the 2004 Genomics Track, making it the track with the most participants in all of TREC 2004.

The data are currently available to track participants on password-protected Web sites but will be made available to non-TREC participants in early 2005. The version of data released in early 2005 will be updated to correct some minor errors associated with the official TREC 2004 data.

3. Ad Hoc Retrieval Task

The goal of the ad hoc task was to mimic conventional searching. The use case was a scientist with a specific information need, searching the MEDLINE bibliographic database to find relevant articles to retrieve.

3.1 Documents

The document collection for the ad hoc retrieval task was a 10-year subset of MEDLINE. We contemplated the use of full-text documents in this task but were unable to procure an adequate amount to represent real-world searching. As such, we chose to use MEDLINE. As noted above, however, despite the widespread availability of on-line, full-text scientific journals at present, most searchers of the biomedical literature still use MEDLINE as an entry point. Consequently, there is great value in being able to search MEDLINE effectively.

The subset of MEDLINE used for the track consisted of 10 years of completed citations from the database inclusive from 1994 to 2003. Records were extracted using the Date Completed (DCOM) field for all references in the range of 19940101 - 20031231. This provided a total of 4,591,008 records. We used the DCOM field and not the Date Published (DP). As a result, some records were published but not completed prior to 1994, i.e., the collection had:

- 2,814 (0.06%) DPs prior to 1980
- 8,388 (0.18%) DPs prior to 1990
- 138,384 (3.01%) DPs prior to 1994

The remaining 4,452,624 (96.99%) DPs were within the 10 year period of 1994-2004.

The data was made available in two formats:

- MEDLINE - the standard NLM format in ASCII text with fields indicated and delimited by 2-4 character abbreviations (uncompressed - 9,587,370,116 bytes, gzipped - 2,797,589,659 bytes)
- XML - the newer NLM XML format (uncompressed - 20,567,278,551 bytes, gzipped - 3,030,576,659 bytes)

3.2 Topics

The topics for the ad hoc retrieval task were developed from the information needs of real biologists and modified as little as possible to create needs statements with a reasonable estimated amount of relevant articles (i.e., more than zero but less than one thousand). The information needs capture began with interviews by 12 volunteers who sought biologists in their local environments. A total of 43 interviews yielded 74 information needs. Some of these volunteers, as well as an additional four individuals, created topics in the proposed format from the original interview data. We aimed to have each information need reviewed more than once but were only able to do this with some, ending up with a total of 91 draft topics. The same individuals then were assigned different draft topics for searching on PubMed so they could be modified to generate final topics with a reasonable number of relevant articles. The track chair made one last pass to make the formatting consistent and extract the 50 that seemed most suitable as topics for the track.

The topics were formatted in XML and had the following fields:

- ID - 1 to 50
- Title - abbreviated statement of information need
- Information need - full statement information need
- Context - background information to place information need in context

We created an additional five "sample" topics, one of which is displayed in Figure 2.

```
<TOPIC>
 <ID>51</ID>
 <TITLE>pBR322 used as a gene vector</TITLE>
 <NEED>Find information about base sequences and restriction maps in plasmids that are used
        as gene vectors.</NEED>
 <CONTEXT>The researcher would like to manipulate the plasmid by removing a particular
        gene and needs the original base sequence or restriction map information of the
        plasmid.</CONTEXT>
 </TOPIC>
```

Figure 2 - Sample topic for ad hoc retrieval task.

3.3 Relevance Judgments

Relevance judgments were done using the conventional "pooling method" whereby a fixed number of top-ranking documents from each official run were pooled and provided to an individual (blinded to the number of groups who retrieved the document and what their search statements were). The relevance assessor then judged each document for the specific topic query as definitely relevant (DR), possibly relevant (PR), or not relevant (NR). A subset of documents were also judged in duplicate to assess interjudge reliability using the kappa measure [5]. For the official results, which required binary relevance judgments, documents that were rated DR or PR were considered relevant.

The pools were built as follows. Each of the 27 groups designated a top-precedence run that would be used for relevance judgments, typically what they thought would be their best-performing run. We took, on average, the top 75 documents for each topic from these 27 runs and eliminated the duplicates to create a single pool for each topic. The average pool size (average number of documents judged per topic) was 976, with a range of 476-1450.

The judgments were done by two individuals with backgrounds in biology. One was a PhD biologist and the other an undergraduate biology student. Table 1 shows the pool size and number of relevant documents for each topic. (It also shows the overall results, to be described later.)

For the kappa measurements, we selected every tenth article from six topics. As each judge had already judged the documents for three of the topics, we compared these extra judgments with the regular ones done by the other judge. The results of the duplicate judgments are shown in Table 2. The resulting kappa score was 0.51, indicating a "fair" level of agreement but not being too different from similar relevance judgment activities in other domains, e.g., [6]. In general, the PhD biologist assigned more articles in the relevant category than the undergraduate.

3.4 Evaluation Measures

The primary evaluation measure for the task was mean average precision (MAP). Results were calculated using the trec_eval program, a standard scoring system for TREC. A statistical analysis was performed using a repeated measures analysis of variance, with posthoc Tukey tests for pairwise comparisons. In addition to analyzing MAP, we also assessed precision at 10 and 100 documents.

3.5 Results

The results of all participating groups are shown in Table 3. The statistical analysis for MAP demonstrated significance across all the runs, with the pairwise significance for the top run (pllsgen4a2) not obtained until the run RMITa about one-quarter of the way down the results.

The best official run was achieved by Patolis Corp. [7]. This run used a combination of Okapi weighting (BM25 for term frequency but with standard inverse document frequency), Porter stemming, expansion of symbols by LocusLink and MeSH records, blind relevance feedback (also known as blind query expansion), and use of all three fields in the query. This group also reported a post-submission run that added the language modeling technique of Dirichlet-Prior smoothing to achieve an even higher MAP of 0.4264.

Table 1 - Ad hoc retrieval topics, number of relevant documents, and average results for all runs.

Topic	Pool	Definitely Relevant	Possibly Relevant	Not Relevant	D & P Relevant	MAP average	P@10 average	P@100 average
1	879	38	41	800	79	0.3073	0.7383	0.2891
2	1264	40	61	1163	101	0.0579	0.2787	0.1166
3	1189	149	32	1008	181	0.0950	0.3298	0.2040
4	1170	12	18	1140	30	0.0298	0.0894	0.0360
5	1171	5	19	1147	24	0.0564	0.1340	0.0349
6	787	41	53	693	94	0.3993	0.8468	0.3938
7	730	56	59	615	115	0.2006	0.4936	0.2704
8	938	76	85	777	161	0.0975	0.3872	0.2094
9	593	103	12	478	115	0.6114	0.7957	0.6196
10	1126	3	1	1122	4	0.5811	0.2532	0.0277
11	742	87	24	631	111	0.3269	0.5894	0.3843
12	810	166	90	554	256	0.4225	0.7234	0.5866
13	1118	5	19	1094	24	0.0288	0.1021	0.0274
14	948	13	8	927	21	0.0479	0.0894	0.0270
15	1111	50	40	1021	90	0.1388	0.2915	0.1800
16	1078	94	53	931	147	0.1926	0.4489	0.2883
17	1150	2	1	1147	3	0.0885	0.0511	0.0115
18	1392	0	1	1391	1	0.6254	0.0660	0.0072
19	1135	0	1	1134	1	0.1594	0.0362	0.0062
20	814	55	61	698	116	0.1466	0.3957	0.2238
21	676	26	54	596	80	0.2671	0.4702	0.2796
22	1085	125	85	875	210	0.1354	0.4234	0.2709
23	915	137	21	757	158	0.1835	0.3745	0.2747
24	952	7	19	926	26	0.5970	0.7468	0.1685
25	1142	6	26	1110	32	0.0331	0.1000	0.0330
26	792	35	12	745	47	0.4401	0.7298	0.2411
27	755	19	10	726	29	0.2640	0.4319	0.1355
28	836	6	7	823	13	0.2031	0.2532	0.0643
29	756	33	10	713	43	0.1352	0.1809	0.1515
30	1082	101	64	917	165	0.2116	0.4872	0.3113
31	877	0	138	739	138	0.0956	0.2489	0.2072
32	1107	441	55	611	496	0.1804	0.6085	0.4787
33	812	30	34	748	64	0.1396	0.2234	0.1647
34	778	1	30	747	31	0.0644	0.0830	0.0668
35	717	253	18	446	271	0.3481	0.8213	0.6528
36	676	164	90	422	254	0.4887	0.7638	0.6700
37	476	138	11	327	149	0.5345	0.7426	0.6564
38	1165	334	89	742	423	0.1400	0.5915	0.4043
39	1350	146	171	1033	317	0.0984	0.3936	0.2689
40	1168	134	143	891	277	0.1080	0.3936	0.2796
41	880	333	249	298	582	0.3356	0.6766	0.6521
42	1005	191	506	308	697	0.1587	0.6596	0.5702
43	739	25	170	544	195	0.1185	0.6915	0.2553
44	1224	485	164	575	649	0.1323	0.6149	0.4632
45	1139	108	48	983	156	0.0286	0.1574	0.0711
46	742	111	86	545	197	0.2630	0.7362	0.4981
47	1450	81	284	1085	365	0.0673	0.3149	0.2355
48	1121	53	102	966	155	0.1712	0.4021	0.2557
49	1100	32	41	1027	73	0.2279	0.5404	0.2049
50	1091	79	223	789	302	0.0731	0.3447	0.2534
Mean	975.1	92.6	72.8	809.7	165.4	0.2171	0.4269	0.2637
Median	978.5	54	44.5	783	115.5	0.1590	0.3989	0.2472
Min	476	0	1	298	1	0.0286	0.0362	0.0062
Max	1450	485	506	1391	697	0.6254	0.8468	0.6700

Table 2 - Kappa results for interjudge agreement in relevant judgments for ad hoc retrieval task.

Judge 2	Definitely relevant	Possibly relevant	Not relevant	Total
Judge 1				
Definitely relevant	62	35	8	105
Possibly relevant	11	11	5	27
Not relevant	14	57	456	527
Total	87	103	469	659

The next best run was achieved by the University of Waterloo [8]. This group used a variety of approaches including Okapi weighting, blind relevance feedback, and various forms of domain-specific query expansion. Their blind relevance feedback made use of usual document feedback as well as feedback from passages. Their domain-specific query expansion included expanding lexical variants as well as expanding acronym, gene, and protein name synonyms.

A number of groups used boosting of word weights in queries or documents. Tsinghua University boosted words in titles and abstracts, along with using blind query expansion [9]. Alias-i Corp. boosted query words in the title and need statements [10]. University of Tampere found value in identifying and using bi-gram phrases [11].

A number of groups implemented techniques, however, that were detrimental. This is evidenced by the OHSU runs, which used the Lucene system "out of the box" that applies TF*IDF weighting [12]. Approaches that attempted to map to controlled vocabulary terms did not fare as well, such as Indiana University [13], University of California Berkeley [14], and the National Library of Medicine [15]. Many groups tried a variety of approaches, beneficial or otherwise, but usually without comparing common baseline or running exhaustive experiments, making it difficult to discern exactly which techniques provided benefit. Figure 3 shows the official results graphically with annotations for the first run statistically significant from the top run as well as the OHSU "baseline."

As typically occurs in TREC ad hoc runs, there was a great deal of variation within individual topics, as is seen in Table 1. Figure 4 shows the average MAP across groups for each topic. Figure 5 presents the same data sorted to give a better indication of the variation across topics. There was a fairly strong relationship between the average and maximum MAP for each topic (Figure 6), while the number of relevant per topic versus MAP was less associated (Figure 7).

4. Categorization Task

In the categorization task, we simulated two of the classification activities carried out by human annotators for the MGI system: a triage task and two simplified variations of MGI's annotation task. Systems were required to classify full-text documents from a two-year span (2002-2003) of three journals, with the first year's (2002) documents comprising the training data and the second year's (2003) documents making up the test data.

One of the goals of MGI is to provide structured, coded annotation of gene function from the biological literature. Human curators identify genes and assign GO codes about gene function with another code describing the type of experimental evidence supporting assignment of the GO code. The huge amount of literature requiring curation creates a challenge for MGI, as their resources are not unlimited. As such, they employ a three-step process to identify the papers most likely to describe gene function:

1. About mouse - The first step is to identify articles about mouse genomics biology. The full text of articles from several hundred journals are searched for the words *mouse*, *mice*, or *murine*. Articles passing this step are further analyzed for inclusion in MGI. At present, articles are searched in a Web browser one at a time because full-text searching is not available for all of the journals included in MGI.

Table 3 - Ad hoc retrieval results, sorted by mean average precision.

Run	Group (reference)	Manual/ Automatic	Mean Average Precision	Relevant at 10 documents	Relevant at 100 documents
pllsgen4a2	patolis.fujita [7]	A	0.4075	6.04	41.96
uwmtDg04tn	u.waterloo.clarke [8]	A	0.3867	6.24	42.1
pllsgen4a1	patolis.fujita [7]	A	0.3689	5.7	39.36
THUIRgen01	tsinghua.ma [9]	M	0.3435	5.82	39.24
THUIRgen02	tsinghua.ma [9]	A	0.3434	5.94	39.44
utaauto	u.tampere [11]	A	0.3324	5.02	32.26
uwmtDg04n	u.waterloo.clarke [8]	A	0.3318	5.68	36.84
PSE	german.u.cairo [18]	A	0.3308	5.86	36.66
tnog3	tno.kraaij [19]	A	0.3247	5.6	36.56
tnog2	tno.kraaij [19]	A	0.3196	5.62	36.04
utamanu	u.tampere [11]	M	0.3128	6.52	38.88
aliasiBase	alias-i [10]	A	0.3094	5.38	34.58
ConversManu	converspeech [20]	M	0.2931	5.82	37.18
RMITa	rmit.scholer [21]	A	0.2796	5.12	31.4
aliasiTerms	alias-i [10]	A	0.2656	4.8	30.3
akoike	u.tokyo (none)	M	0.2427	4.48	31.3
OHSUNeeds	ohsu.hersh [12]	A	0.2343	3.84	26.46
tgnSplit	tarragon [22]	A	0.2319	4.86	29.26
UIowaGN1	u.iowa [23]	A	0.2316	4.76	28.5
tq0	nlm.umd.ul [15]	A	0.2277	5.12	30.1
OHSUAll	ohsu.hersh [12]	A	0.2272	4.32	27.76
LHCUMDSE	nlm.umd.ul [15]	A	0.2191	3.9	24.18
akoyama	u.tokyo (none)	M	0.2155	4.52	25.62
PDTNsmp4	u.padova [24]	A	0.2074	4.56	23.18
PD50501	u.padova [24]	A	0.2059	4.42	25.18
RMITb	rmit.scholer [21]	A	0.2059	4.56	27.26
UBgtNormJM1	suny.buffalo [25]	A	0.2043	4.34	25.38
ConversAuto	converspeech [20]	A	0.2013	3.88	22.8
york04g2	york.u [26]	M	0.2011	5.5	25.8
tgnNecaux	tarragon [22]	A	0.1951	4.08	23.58
lga1	indiana.u.seki [13]	A	0.1833	3.08	22.86
york04g1	york.u [26]	A	0.1794	4.14	26.96
lga2	indiana.u.seki [13]	A	0.1754	3.1	20.22
rutgersGAH1	rutgers.dayanik [16]	A	0.1702	4.66	26.76
wdvqlxa1	indiana.u.yang [27]	A	0.1582	4.2	24.78
wdvqlx1	indiana.u.yang [27]	A	0.1569	4.26	24.26
DCUmatn1	dubblincity.u [28]	M	0.1388	3.28	17.84
BioTextAdHoc	u.cberkeley.hearst [14]	A	0.1384	3.76	23.76
shefauto2	u.sheffield.gaizauskas [29]	A	0.1304	3.66	18.5
rutgersGAH2	rutgers.dayanik [16]	A	0.1303	3.42	19.48
shefauto1	u.sheffield.gaizauskas [29]	A	0.1294	3.54	18.92
run1	utwente (none)	M	0.1176	1.5	10.5
MeijiHilG	meiji.u [30]	A	0.0924	2.1	15.24
DCUma	dubblincity.u [28]	M	0.0895	2.4	15.46
csusm	u.sanmarcos [31]	M	0.0123	0.44	1.6
edinauto2	u.edinburgh.sinclair [32]	A	0.0017	0.46	1.6
edinauto5	u.edinburgh.sinclair [32]	A	0.0012	0.36	1.3
Mean			0.2074	4.48	26.46

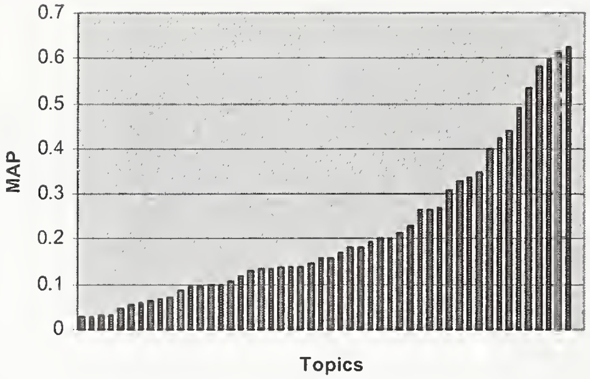


Figure 5 - MAP by topic for the ad hoc task sorted by MAP.

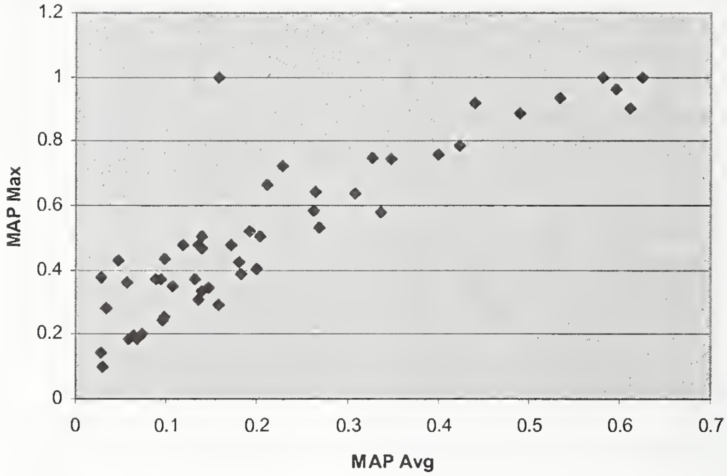


Figure 6 - The maximum MAP plotted vs. average MAP for the ad hoc retrieval task runs.

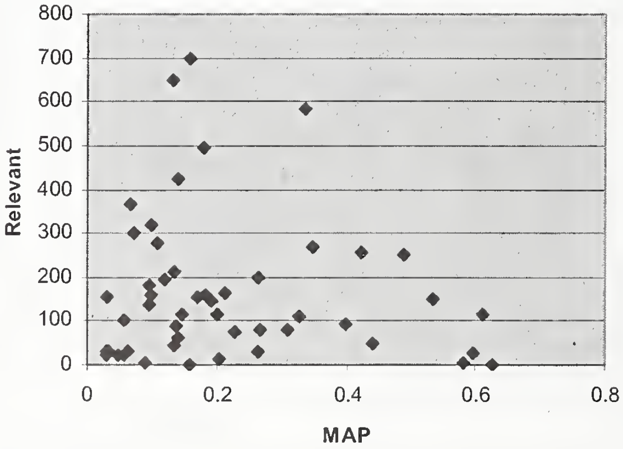
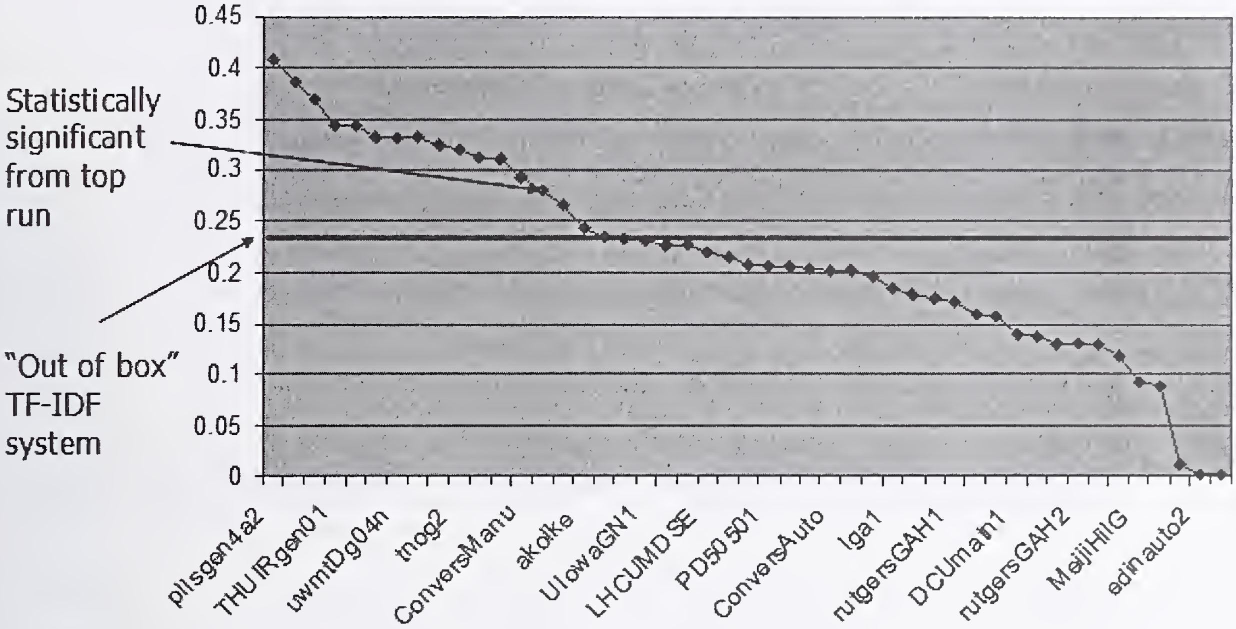


Figure 7 - The number of relevant per topic plotted vs. MAP for the ad hoc retrieval task.



Statistically significant from top run

"Out of box" TF-IDF system

Figure 3 - Ad hoc retrieval runs sorted by MAP score. The highest run to obtain statistical significance (RMITa) from the top run (pllsgen4a2) is denoted, along with the "out of the box" TF*IDF run (OHSUNeeds) are annotated.

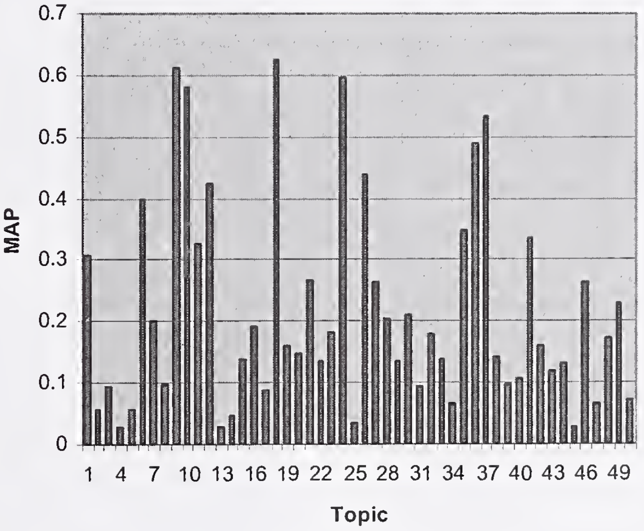


Figure 4 - MAP by topic for the ad hoc task.

2. Triage - The second step is to determine whether the identified articles should be sent for curation. MGI curates articles not only for GO terms, but also for other aspects of biology, such as gene mapping, gene expression data, phenotype description, and more. The goal of this triage process is to limit the number of articles sent to human curators for more exhaustive analysis. Articles that pass this step go into the MGI system with a tag for GO, mapping, expression, etc.. The rest of the articles do not go into MGI. Our triage task involved correctly classifying which documents had been selected for GO annotation in this process.
3. Annotation - The third step is the actual curation with GO terms. Curators identify genes for which there is experimental evidence to warrant assignment of GO codes. Those GO codes are assigned, along with a code for each indicating the type of experimental evidence. There can more than one gene assigned GO codes in a given paper and there can be more than one GO code assigned to a gene. In general, and in our collection, there is only one evidence code per GO code assignment per paper. Our annotation task involved a modification of this annotation step as described below.

4.1 Documents

The documents for the categorization task consisted of articles from three journals over two years, reflecting the full-text documents we were able to obtain from Highwire Press (www.highwire.org). Highwire is a "value added" electronic publisher of scientific journals. Most journals in their collection are published by professional associations, with the copyright remaining with the associations. Highwire originally began with biomedical journals, but in recent years has expanded into other disciplines. They have also supported IR and related research by acting as an intermediary between consenting publishers and information systems research groups who want to use their journals, such as the Genomics Track.

The journals available and used by our track this year were *Journal of Biological Chemistry* (JBC), *Journal of Cell Biology* (JCB), and *Proceedings of the National Academy of Science* (PNAS). These journals have a good proportion of mouse genome articles. Each of the papers from these journals was provided in SGML format based on Highwire's Document Type Definition (DTD). We used articles from the year 2002 for training data and from 2003 for test data. The documents for the categorization tasks came from a subset of articles having the words *mouse*, *mice* or *murine* as described above. We created a crosswalk file (look-up table) that matched an identifier for each Highwire article (its file name) and its corresponding PubMed ID (PMID). Table 4 shows the total number of articles in each journal and the number in each journal included in subset used by the track. The SGML training document collection was 150 megabytes in size compressed and 449 megabytes uncompressed. The SGML test document collection was 140 megabytes compressed and 397 megabytes uncompressed.

Since MGI annotation lags behind article publication, a not insubstantial number of papers have been selected for annotation but not yet annotated. From the standpoint of the triage subtask, we wanted to use all of these articles as positive examples, since they all were selected for GO annotation. However, we could not use the articles not yet annotated for the annotation hierarchy task, since we did not have the annotations. We also needed a set of negative examples for the annotation hierarchy task and chose to use articles selected for action by MGI for other (i.e., non-GO annotation) actions. Figure 8 shows the groups of documents and how they were assigned into being positive and negative examples for the subtasks.

4.2 Triage Subtask

The goal of this task was to correctly identify papers that were deemed to have experimental evidence warranting annotation with GO codes. Positive examples included papers designated for GO annotation by MGI. As noted above, some of these papers had not yet been annotated. Negative examples were all papers not designated for GO annotation in the operational MGI system. For the training data (2002), there were 375 positive examples, meaning that there were 5837-375 = 5462 negative examples. For the test data (2003), there were 420 positive examples, meaning that there were 6043-420 = 5623 negative examples. It should also be noted that the MGI system is, like most operational databases, continuously updated, so the data for the track represented a snapshot of the database obtained in May, 2004. (As described later, an updated version of the data will be available in 2005.)

Table 4 - Number of papers total and available in the *mouse*, *mus*, or *murine* subset.

Journal	**2002 papers - total, subset**	**2003 papers - total, subset**	**Total papers - total, subset**
JBC	6566, 4199	6593, 4282	13159, 8481
JCB	530, 256	715, 359	1245, 615
PNAS	3041, 1382	2888, 1402	5929, 2784
Total papers	10137, 5837	10196, 6043	20333, 11880

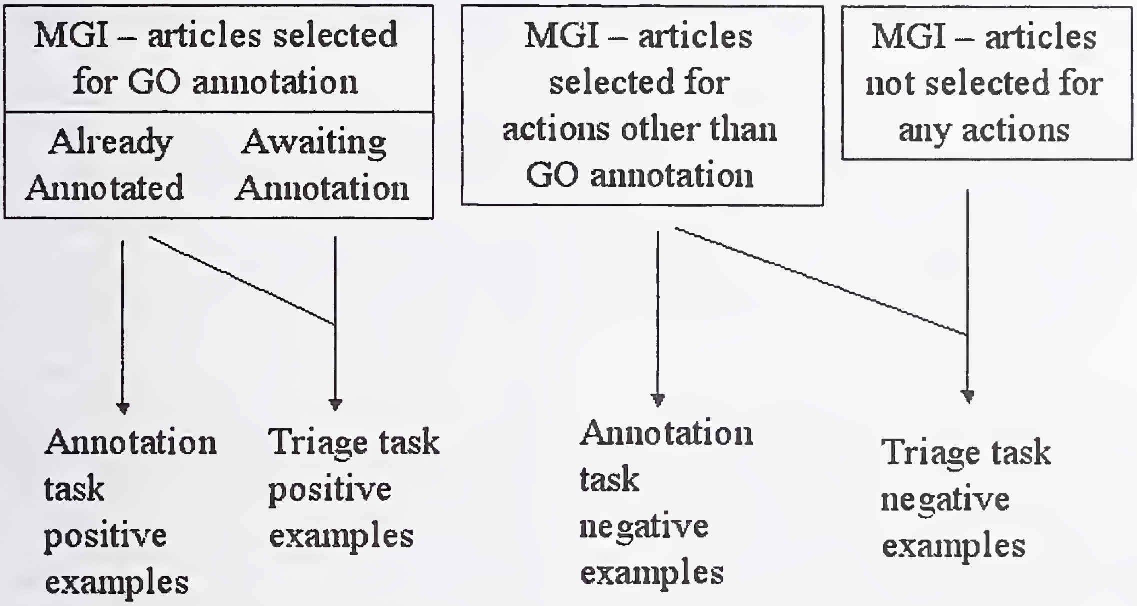

MGI – articles selected for GO annotation
Already Annotated
Awaiting Annotation
MGI – articles selected for actions other than GO annotation
MGI – articles not selected for any actions
Annotation task positive examples
Triage task positive examples
Annotation task negative examples
Triage task negative examples

Figure 8 - Grouping of documents for categorization subtasks.

The evaluation measure for the triage task was the utility measure often applied in text categorization research and used by the former TREC Filtering Track. This measure contains coefficients for the utility of retrieving a relevant and retrieving a nonrelevant document. We used a version that was normalized by the best possible score:

$$U_{norm} = U_{raw} / U_{max}$$

where U_{norm} was the normalized score, U_{raw} the raw score, and U_{max} the best possible score.

The coefficients for the utility measure were derived as follows. For a test collection of documents to categorize, U_{raw} is calculated as:

$U_{raw} = (u_r * \text{relevant-docs-retrieved}) + (u_{nr} * \text{nonrelevant-docs-retrieved})$

where:

- u_r = relative utility of relevant document
- u_{nr} = relative utility of nonrelevant document

We used values for u_r and u_{nr} that were driven by boundary cases for different results. In particular, we wanted (thought it was important) the measure to have the following characteristics:

- Completely perfect prediction - $U_{norm} = 1$
- All documents designated positive (triage everything) - $1 > U_{norm} > 0$
- All documents designated negative (triage nothing) - $U_{norm} = 0$
- Completely imperfect prediction - $U_{norm} < 0$

In order to achieve the above boundary cases, we had to set $u_r > 1$. The ideal approach would have been to interview MGI curators and use decision-theoretic approaches to determine their utility. However, time constraints did not allow this. Deciding that the triage-everything approach should have a higher score than the triage-nothing approach, we estimated that a U_{norm} in the range of 0.25-0.3 for the triage-everything condition would be appropriate. Solving for the above boundary cases with U_{norm} ~ 0.25-0.3 for that case, we obtained a value for u_r ~ 20. To keep calculations simple, we choose a value of u_r = 20. Table 5 shows the value of U_{norm} for the boundary cases.

The measure U_{max} was calculated by assuming all relevant documents were retrieved and no nonrelevant documents were retrieved, i.e., $U_{max} = u_r$ * all-relevant-docs-retrieved.

Thus, for the training data,

U_{raw} = (20 * relevant-docs-retrieved) - nonrelevant-docs-retrieved

U_{max} = 20 * 375 = 7500

U_{norm} = [(20 * relevant-docs-retrieved) - nonrelevant-docs-retrieved] / 7500

Likewise, for the test data,

U_{raw} = (20 * relevant-docs-retrieved) - nonrelevant-docs-retrieved

U_{max} = 20 * 420 = 8400

U_{norm} = [(20 * relevant-docs-retrieved) - nonrelevant-docs-retrieved] / 8400

The results of the triage subtask are shown in Table 6. A variety of groups used classifiers based on machine learning techniques. The higher scoring runs tended to make use of MeSH terms in some fashion. The best performing run came from Rutgers University, using the MEDLINE record, weighting, and filtering by the MeSH term *Mice* [16]. They achieved a U_{norm} of 0.6512. However, this group also noted that the MeSH term *Mice* alone scored better than all but the single top run, with a U_{norm} of 0.6404. This meant that no other approach was better able to classify documents for triage than simply using the MeSH term *Mice* from the MEDLINE record. Of course, this run only achieved a recall of about 15% (with a recall of 89%), so this feature is far from a perfect predictor. In an another analysis of the data, Cohen noted that there was conceptual drift across the collection, with the features identified as strong predictors in the training data not necessarily continuing to be strong predictors in the test data [12]. All of the triage subtask results are shown graphically in Figure 9, along with the utility for the MeSH term *Mice* and the decision to select all articles.

4.3 Annotation Subtask

The primary goal of this task was, given an article and gene name, to correctly identify which of the GO hierarchies (also called domains) had terms within them that were annotated by the MGI curators. Note that the goal of this task was not to select the actual GO term, but rather to select the one or more GO hierarchies (molecular function, biological process, or cellular component) from which terms had been selected to annotate the gene for the article. Papers that were annotated had terms from one to three hierarchies.

For negative examples, we used 555 papers that had a gene name assigned but were used for other purposes by MGI. As such, these papers had no GO annotations. These papers did, however, have one or more gene assigned by MGI for the other annotation purposes.

A secondary subtask was to identify the correct GO evidence code that went with the hierarchy code. Only two groups took part in this subtask.

Table 7 shows the contents and counts of the data files for this subtask. For the training data, there were a total of 504 documents that were either positive (one or more GO terms assigned) or negative (no GO terms assigned) examples. From these documents, a total of 1291 genes had been assigned by MGI. (The Genes file contained the MGI identifier, the gene symbol, and the gene name. It did not contain any other synonyms.) There were 1418 unique possible document-gene pairs in the training data. The data from the first three rows of Table 7 differ from the rest in that they contained data merged from positive and negative examples. These were what would be used as input for systems to nominate GO domains or the GO domains plus their evidence codes per the annotation task. When the test data were released, these three files were the only ones that were provided.

For the positive examples in the training data, there were 178 documents and 346 document-gene pairs. There were 589 document-gene name-GO domain tuples (out of a possible 346 * 3 = 1038). There were 640 document-gene name-GO domain-evidence code tuples. A total of 872 GO plus evidence codes had been assigned to these documents. For the negative examples, there were 326 documents and 1072 document-gene pairs. This meant that systems could possibly assign 1072*3 = 3216 document-gene name-GO domain tuples.

HARD Track Overview in TREC 2004
High Accuracy Retrieval from Documents

James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst

Abstract

The HARD track of TREC 2004 aims to improve the accuracy of information retrieval through the use of three techniques: (1) query metadata that better describes the information need, (2) focused and time-limited interaction with the searcher through "clarification forms", and (3) incorporation of passage-level relevance judgments and retrieval. Participation in all three aspects of the track was excellent this year with about 10 groups trying something in each area. No group was able to achieve huge gains in effectiveness using these techniques, but some improvements were found and enthusiasm for the clarification forms (in particular) remains high. The track will run again in TREC 2005.

1 Introduction

The High Accuracy Retrieval from Documents (HARD) track explores methods for improving the accuracy of document retrieval systems. It does so by considering three questions:

1. Can additional metadata about the query, the searcher, or the context of the search provide more focused and therefore accurate results? These metadata items generally do not directly affect whether or not a document is on topic, but they do affect whether it is relevant. For example, a person looking for introductory material will not find an on-topic but highly technical document relevant.

2. Can highly focused, short-duration, interaction with the searcher be used to improve the accuracy of a system? Participants created "clarification forms" generated in response to a query—and leveraging any information available in the corpus—that were filled out by the searcher. Typical clarification questions might ask whether some titles seem relevant, whether some words or names are on topic, or whether a short passage of text is related.

3. Can passage retrieval be used to effectively focus attention on relevant material, increasing accuracy by eliminating unwanted text in an otherwise useful document? For this aspect of the problem, there are challenges in finding relevant passages, but also in determining how best to evaluate the results.

The HARD track ran for the second time in TREC 2004. It used a new corpus and a new set of 50 topics for evaluation. All topics included metadata information and clarification forms were considered for each of them. Because of the expense of sub-document relevance judging, only half of the topics were used in the passage-level evaluation.

A total of 16 sites participated in HARD, up from 14 the year before. Interest remains strong, so the HARD track will run again in TREC 2005, but because of funding uncertainties will only address a subset of the issues. Exactly what is included and how it takes place will be determined by interested participants. Information about the track will be available at the track's Web page, `http://ciir.cs.umass.edu/research/hard` (the contents of the site are not predictable after 2005).

Topic creation, clarification form entry, and relevance judging were all carried out by the Linguistic Data Consortium (LDC) at the University of Pennsylvania (http://www.ldc.upenn.edu). The annotation work was supported in part by the DARPA TIDES project.

Evaluation of runs using the judgments from the LDC was carried out by NIST.

The remainder of this document discusses the HARD 2004 track and provides an overview of some of its results. Additional details on results are available in the TREC papers from the participating sites.

2 HARD Corpus

The HARD 2004 evaluation corpus itself consisted entirely of English text from 2003, most of which is newswire. The specific sources and approximate amounts of material are:

Source	Abbrev	Num docs	Size (Mbs)
Agence France Press	AFP	226,777	497
Associated Press	APW	236,735	644
Central News Agency	CNA	4,011	6
LA Times/Wash Post	LAT	34,145	107
New York Times	NYT	27,835	105
Salon.com	SLN	3,134	28
Ummah Press	UMM	2,557	5
Xinhua (English)	XIN	117,516	183
Totals		652,710	1,575

This information was made available to participating sites with a research license. The data was provided free of charge, though sites interested in retaining the data after the HARD track ended were required to make arrangements with the LDC to do so.

3 Topics

Topics were an extension of typical TREC topics: they included (1) a statement of the topic and (2) a description of metadata that a document must satisfy to be relevant, even if it is on topic. The topics were represented in XML and included the following components:

- *number* is the topic's number–e.g., HARD-003.
- *title* is a short, few word description of the topic.
- *description* is a sentence-length description of the topic.
- *topic-narrative* is a paragraph-length description of the topic. This component did not contain any mention of metadata restrictions. It is intended purely to define what is "on topic."
- *metadata-narrative* is a topic author's description of how metadata is intended to be used. This description helps make it clear how the topic and metadata were intended to interact.
- *retrieval-element* indicates whether the judgments (hence retrieval) should be at the *document* or *passage* level. For HARD 2004, half of the topics were annotated at the passage level.
- The following metadata fields were provided:
  - *familiarity* had a value of *little* or *much*. It affected whether a document was relevant, but not whether it was on topic.
  - *genre* had values of *news-report*, *opinion-editorial*, *other*, or *any*. It affected whether a document was relevant, but not whether it was on topic.
  - *geography* had values of *US*, *non-US*, or *any*. It affected whether a document was relevant, but not whether it was on topic.
  - *subject* describes the subject domain of the topic. It is a free-text field, though the LDC attempted to be consistent in the descriptions it used. It affected whether or not a document was on-topic.

- *related-text.on-topic* provided an example of text that the topic's author considered to be on-topic but not relevant.
- *related-text.relevant* provided an example of text that the topic's author considered to be relevant (and therefore also on-topic).

During topic creation, the LDC made an effort to have topics vary across each of the indicated metadata items.

The following is a sample topic from the evaluation corpus (topic HARD-428). Some particularly long sections of the topic have been elided.

```
<topic>

<number>
HARD-428
</number>

<title>
International organ traffickers
</title>

<description>
Who creates the demands in the international ring of organ trafficking?
</description>

<topic-narrative>
Many countries are institutionalizing legal measures to prevent the
selling and buying of human organs. Who, in the ring of international
organ trafficking, are the "buyers" of human organs? Any information
that identifies 'where' they are or 'who' they may be will be
considered on topic; the specificity of info does not matter. Also,
the story must be about international trafficking. Stories that only
contain information about the "sellers" of organs or those that focus
on national trafficking will be off topic.
</topic-narrative>

<metadata-narrative>
Subject (CURRENT EVENTS) is chosen as it is expected that such
articles will have more information about the identities of the
parties involved. Genre (NEWS) is expected to exclude stories that
tends to focus on ethical matters.
</metadata-narrative>

<retrieval-element>
passage
</retrieval-element>

<metadata>
   <familiarity>
   little
   </familiarity>

   <genre>
   news-report
   </genre>
```

```
    <geography>
    any
    </geography>

    <related-text>
      <on-topic>
      Every day, 17 Americans die of organ failure. In Israel, the average
      wait for a kidney transplant is four years. In response, a global gray
      market has bloomed. In India, for example, poor sellers are quickly...
      </on-topic>

      <relevant>
      At least 30 Brazilians have sold their kidneys to an international
      human organ trafficking ring for transplants performed in South
      Africa, with Israel providing most of the funding, says a legislative...
      </relevant>
    </related-text>

    <subject>
    CURRENT EVENTS
    </subject>
</metadata>
</topic>
```

4 Relevance judgments

For each topic, documents that are annotated get one of the following judgments:

- OFF-TOPIC means that the document does not match the topic. (As is common in TREC, a document without any judgment is assumed to be off topic for evaluation purposes.)
- ON-TOPIC means that the document does match the topic but that it does not satisfy the provided metadata restrictions. Given the metadata items listed above, that means it either does not satisfy the FAMILIARITY, GENRE, or GEOGRAPHY items (note that SUBJECT affects whether a story is on topic).
- RELEVANT means that the document is on topic *and* it satisfies the appropriate metadata.

In addition, if the *retrieval element* field is *passage* then each judgment comes with information that specifies which portions of the documents are relevant.

To specify passages, HARD used the same approach used by the question answering track [Voorhees, 2005]. A passage is specified by its byte offset and length. The offset will be from the "<" in the "<DOC>" tag of the original document (an offset of zero would mean include the "<" character). The length will indicate the number of bytes that are included. If a document contains multiple relevant passages, the document will be listed multiple times.

The HARD track used the standard TREC pooling approach to find possible relevant documents. The top 85 documents from one baseline and one final run from each submitted system were pooled (i.e., 85 times 16 times 2 documents). The LDC considered each of those documents as possibly relevant to the topic.

Across all topics, the LDC annotated 36,938 documents, finding 3,026 that were on topic and relevant and another 744 that were on topic but not relevant. Topics ranged from one on topic and relevant document to 519; from 1 on topic but not relevant document to 70.

5 Training data

The LDC provided 20 training topics and 100 judged documents per topic. The topics incorporated a selection of metadata values and came with relevance judgments.

In addition, the LDC provided a mechanism to allow sites to validate their clarification forms. Sites could send a form to the LDC and get back confirmation that the form was viewable and some "random" completion of the form. The resulting information was sent back to the site in the same format that was used in the evaluation. (No one took advantage of such a capability.)

6 Clarification forms

A unique aspect of the HARD track is that it provides access to the person who formulated the query and will be doing the annotation. It allows sites to get a small amount of additional information from that person by providing a small Web page as a form with clarification questions, check boxes, etc. for the searcher to fill in.

The assessor spent no more than three (3) minutes filling out the form for a particular topic. If some portions of a form were not filled out when the time expired, those portions were left blank. Sites were aware of the time limit and were encouraged to keep their forms small—however, several (perhaps most) sites built longer forms intending to get whatever they could within three minutes rather than building forms designed to be filled in quickly.

In order to avoid implementation issues, systems were required to restrict the forms to simple HTML without Javascript, images, and so on. They were also told what would be the hardware configuration used by annotators, so they could tailor the presentation appropriately if desired.

The LDC reported that the annotators enjoyed filling out clarification forms immensely—if only because it was an entirely new type of annotation task for them.

7 Results format

Results were returned for evaluation in standard TREC format extended, though, to support passage-level submissions since it possible that the searcher's preferred response is the best passage (or sentence or phrase) of relevant documents. Results included the top 1000 documents (or top 1000 passages) for each topic, one line per document/passage per topic. Each line had the format:

topic-id Q0 docno rank score tag psg-offset psg-length

where:

- *topic-id* represents the topic number from the topic (e.g., HARD-001)
- *"Q0"* is a constant provided for historical reasons
- *docno* represents the document that is being retrieved (or from which the passage is taken)
- *rank* is the rank number of the document/passage in the list. Rank should start with 1 for the document/passage that the system believes is most likely to be relevant and continue to 1000.
- *score* is a system-internal score that was assigned to the document/passages. High values of score are assumed to be better, so score should generally drop in value as rank increases.
- *tag* is a unique identifier for this run by the site.
- *psg-offset* indicates the byte-offset in document docno where the passage starts. A value of zero represents the "<" in "<DOC>" at the start of the document. A value of negative one (-1) means that no passage has been selected and the entire document is being retrieved.
- *psg-length* represents how many bytes of the document are included in the passage. A value of negative one (-1) must be supplied when psg-offset is negative one.

8 Evaluation approach

Results were evaluated at the document level, both in light of (HARD) and ignoring (SOFT) the query metadata. Ranked lists were also evaluated incorporating passage-level judgments. We discuss each evaluation in this section.

Five of the 50 HARD topics (401, 403, 433, 435, and 450) had no relevant (*and* on topic) documents. That is, although there were documents that matched the topics, no document in the pool matched the topic *and* the query metadata. Accordingly, those five topics were dropped from both the HARD and SOFT evaluations. (They could have been kept for the SOFT evaluation, but then the scores of the two evaluations would not have been comparable.)

8.1 Document-level evaluation

In the absence of passage information, evaluation was done using standard mean average precision. There were two variants, one for HARD judgments and one for SOFT.

Some of the runs evaluated in this portion were actually passage-level runs and could therefore include a document at multiple points in the ranked list—i.e., because more than one passage was considered likely to be relevant. For the document-level evaluation, only the first occurrence of a document in the ranked list was considered. Subsequent occurrences were "deleted" from the ranked list. (That meant that it was possible for a site to submit 1000 items in a ranked list, but have fewer than 1000 documents ranked.)

8.2 Passage-level evaluation

Two passage measures were explored for HARD 2004. The first was the same one used in HARD 2003, passage R-precision. Some research at UMass Amherst demonstrated an extremely strong bias in favor of short passages, so a second measure was also explored.

8.2.1 Passage R-Precision

In a nutshell, this evaluation measure considers the "true" relevant R passages as found by annotators. It considers the top R passage returned by a system and counts the proportion of characters that overlap relevant passages. It incorporates a penalty for repeating text in multiple passages. More details are provided below.

The passage level evaluation for a topic consists of values for passage recall, passage precision, and the F score at cutoff 5, 10, 15, 20, 30, 50, and 100, plus a R-precision score. As with standard document level evaluation, a cutoff is the rank within the result set such that passages at or above the cutoff are "retrieved" and all other passages are not retrieved. So, for example, if the cut-off is 5 the passage recall and precision are computed over the top 5 passages. R-precision is defined similarly to the document level counterpart: it is the passage precision after R passages have been retrieved where R is the number of relevant passages for that topic. We are using passage R-precision as an evaluation measure reported for the track because it is a cutoff-based measure that tracks mean average precision extremely closely in document evaluations.

The following is an operational definition of passage recall and precision as used in the evaluation. For each relevant passage allocate a string representing all of the character positions contained within the relevant passage (i.e., a relevant passage of length 100 has a string of length 100 allocated). Each passage in the retrieved set marks those character positions in the relevant passages that it overlaps with. A character position can be marked at most once, regardless of how many different retrieved passages contain it. (Retrieved passages may overlap, but relevant passages do not overlap.) The passage recall is then defined as the average over all relevant passages of the fraction of the passage that is marked. The passage precision is defined as the total number of marked character positions divided by the total number of characters in the retrieved set. The F score is defined in the same way as for documents, assigning equal weight to recall and precision: F = (2*prec*recall)/(prec+recall) where F is defined to be 0 if prec+recall is 0. We included the F score because set-based recall and precision average extremely poorly but F averages well. R-precision also averages well.

In all of the above, a document is treated as a (potentially long) passage. That is, the relevant "passage" starts at the beginning of the document and is as long as the document. (These are represented in the judgment file as passages with -1 offset and -1 length, but are treated as described above.) For any topic, a retrieved document (i.e., where offset and length are -1) is again just a passage with offset 0 and length the length of the document.

Using the above definition of passage recall, passage recall and standard document level recall are identical when both retrieved and relevant passages are whole documents. That is not true for this definition of passage precision. Passage precision will be greater when a shorter irrelevant document is retrieved as compared to when a longer irrelevant document is retrieved. This makes sense, but is different from standard document level precision.

8.2.2 Passage-level bpref

Some explorations at UMass Amherst showed that passage R-precision could be improved dramatically by splitting existing passages into smaller pieces. For example, by splitting the top-ranked passages into 32 pieces and then using the top R of those (rather than the top R original passages), the value of passage R-precision increased by 128%.

Although numerous measures were considered, a variation of bpref [Buckley and Voorhees, 2004] was finally selected. In this measure, the top 12,000 characters of the system's ranked list of passages was considered (intended to correspond roughly to 10 normal sized passages).

As a document evaluation measure, bpref considers two sets of documents: a relevant set and a non-relevant set. The assumption is that if a document A is taken from the first set and B is taken from the second, then the user has a binary preference that A be ranked higher than B. The measure counts the proportion of times that the user's implied set of preferences is satisfied. A perfect system would rank all known relevant documents above all known non-relevant documents, would thereby satisfy all of the user's preferences, and receive a score of 1.0. The worst possible score is zero, and systems will normally score somewhere in the middle.

To extend this measure to passages, we consider character-level preferences. We assert that all relevant characters should be presented before any non-relevant characters and count the proportion of preferences that are satisfied. Note that the choice of character as the base unit is arbitrary and made for reasons of simplicity. It could have been word, phrase, or even sentence, but each of those would require algorithmic decisions about boundaries between units that are not necessary for character-level decisions. We believe (though have not investigated) that different units will merely change the scale of results.

9 Protocol

The HARD 2004 track ran from May through August of 2004. On June 25th, sites received the 50 evaluation topics, but without any of the metadata fields provided. That is, they received just the title, description, and narrative information, a format consistent with past "ad hoc" TREC tracks.

Using that base information, sites were asked to do their best to rank documents for relevance and return the ranked list of documents (not passages). These were the "baseline runs" and were due to NIST on July 9th.

In addition, sites could optionally generate up to two clarification forms that the LDC annotators would fill out. These forms were due to the LDC on July 16th

On July 29th, the filled-out forms were returned to sites and the metadata fields of the topics were released to all sites, regardless of whether they used clarification forms. Sites could use any of that information to produce improved ranked lists. The final runs, incorporating everything they could, were due to NIST on August 5th.

As described above, one baseline run and one final run were used from each site. The top 85 documents from each of those runs were pooled together and used by the LDC for judging. For topics that required passage-level judgment, the annotator marked passages as relevant as soon as a relevant document was found.

10 Participation

The following 16 sites participated in the HARD track of TREC 2004. The first three columns indicate whether the site used metadata values, clarification forms, or passage retrieval in any of their submitted runs.

Meta	CF	Psgs	Site
Y	Y	Y	Chinese Academy of Science, Institute of Software [Sun et al., 2005]
N	Y	N	Clairvoyance Corporation [Evans et al., 2005]
N	Y	Y	Indiana University [Yang et al., 2005]
N	Y	N	Microsoft Research Cambridge [Zaragoza et al., 2005]
Y	N	N	The Robert Gordon University [Harper et al., 2005]
Y	N	N	Rutgers University [Belkin et al., 2005]
?	?	?	Tsinghua University
Y	N	Y	University of Chicago [Levow, 2005]
?	?	?	University of Cincinnati
N	Y	Y	University of Illinois at Urbana-Champaign [Jiang and Zhai, 2005]
N	Y	Y	University of Maryland & Johns Hopkins University [He et al., 2005]
Y	Y	Y	University of Massachusetts Amherst [Abdul-Jaleel et al., 2005]
N	Y	N	University of North Carolina at Chapel Hill [Kelly et al., 2005]
Y	N	N	University of Twente[Rode and Hiemstra, 2005]
N	Y	Y	University of Waterloo & Bilkent University [Vechtomova and Karamuftuoglu, 2005]
Y	Y	Y	York University [Huang et al., 2005]
7	10	8	COUNTS

(No information was reported for Tsinghua University or the University of Cincinnati, and they did not provide a paper on this track to TREC for publication.)

It is interesting to note the wide range of ways that the different purposes of the track were exploited. Only three sites used all three possible components of the track. The clarification forms were the most popular, but not by a wide margin.

11 Results

This section provides a sketch of some of the results found by participating sites. Further and more detailed information is available in the sites individual papers.

11.1 Use of metadata

For the most part, sites built models for the geography, genre, and subject metadata categories. They typically used text classification techniques to decide whether a document matched the category. Some sites used the Web to collect more data relevant to the category. And some built manual term lists for classification (mostly for geography information).

In general, sites were unable to demonstrate substantial gains in effectiveness using metadata. Since metadata differentiated between relevant and merely on-topic documents, a run using metadata should score much better on "hard" measures (where only relevant documents are counted as relevant) and "soft" measures (where on-topic documents are also counted as relevant). Several runs were able to improve in that direction, though not by huge margins.

Some of these results are because topics tended not to require the metadata to improve performance. For example, *AIDS in Africa* is obviously a non-US topic, and being told that it is not US is of little value.

The University of North Carolina asked (in clarification forms) the user how many times they had searched before for each topic. They then showed that users who had claimed low familiarity in metadata also had not previously searched often for this topic. They did not use the metadata to aid retrieval, but cleverly used the clarification form to show how familiarity metadata could be collected [Kelly et al., 2005].

The University of Waterloo also did not use metadata for retrieval, but did a very nice analysis using the familiarity metadata. Users with low familiarity selected fewer phrases in Waterloo's clarification forms. User's with low familiarity were helped by the clarification forms but users with much familiarity were hurt [Vechtomova and Karamuftuoglu, 2005].

11.2 Use of clarification forms

Clarification forms allowed sites to ask the user anything about the topic that could be expressed in simple HTML. Most requested information asked for judgments on keywords, documents, or passages. One site asked whether presented passages were of about the right length, presumably to get a handle on the right amount of information that should be returned. Several sites included free-form entry of phrases, other keywords, or related text at the end of their clarification forms.

When sites asked for keywords, they had usually found words or phrases that their system suspected were related to the topic. These might be words or phrases appearing in top-ranked documents, synonyms of query words found using Wordnet (for example), extracted noun phrases or named entities, or ranges of time that where relevant material would appear.

Document-style requests generally asked for a judgment of relevant for the passage. That was often the title and a few keywords from a document, the passage most likely to be relevant ("best passage"), or a cluster of documents represented by titles and/or key words. The set of documents, passages, or clusters chosen for presentation were either the top-ranked set or a set modified to incorporate some notion of novelty—i.e., do not present two highly similar documents for judgment.

Clarification forms were very popular, very fun, provided an open ended framework for experimentation, and were by those counts very successful. On the other hand, most sites limited themselves to keyword and text relevance feedback rather than trying more novel techniques, so the "open ended" nature has not (yet) encouraged new ideas.

The value of clarification forms remains elusive to determine. Many sites saw some gains from their clarification forms, but there were several sites that achieved their best performance—or nearly their best—on the baseline runs. Unquestionably work should consider on clarification forms because they are popular, though until more impressive gains are seen, their value will debatable.

11.3 Use of passages

As described in Section 8, two measures for passage retrieval were considered, but others were compared. Two get a sense of how similar they were, we investigated the correlation between bpref at 12,000 characters. (That measure was declared "primary" in the track guidelines, but sufficiently late in the process that some sites fit to the passage R-precision measure.)

- Precision at 12,000 characters measured the proportion of characters that were relevant in the top 12,000 characters. It showed a 99% correlation.

- Character R-precision (similar to passage R-precision, but a character-oriented evaluation where R is the total number of relevant *characters* not passages). It showed an 88% correlation.

- Passage F1 at 30 passages retrieved showed a 90% correlation.

- Passage precision at 10 passages showed an 80% correlation.

- Passage R-precision (last year's official measure) showed a 45% correlation.

If nothing else, these results should suggest that sites training their systems to optimize passage R-precision should not be expected to do well on the character bpref measure.

Passage retrieval systems often use fixed-length passages of some number of words or characters, treating those passages as if they were documents. Some sites tried to generate appropriately sized passages using HMMs, retrieving and then merging highly ranked adjacent sentences, or looking for runs of text where the query terms are highly dense. Most sites scored passages and then combined the passage score with the document score in one way or another.

There was substantially more activity in passage retrieval for HARD 2004 than in 2003. However, the issue of how best to resolve variable-length passage retrieval with variable-length passage "truth" judgments remains open and begs for substantially more exploration. There are clear problems with the passage R-precision measure, but the character bpref is also not without issues. Unfortunately, the HARD 2005 track will be dropping passage retrieval because of funding issues.

11.4 Overall results

When measured by topicality (i.e., when on-topic and/or relevant documents are the target), the top runs were all automatic and used both the title and description. Some top runs used clarification forms, passage retrieval, and the (hard) related text information. A few top runs used the geography and genre metadata fields and a couple used the topic narrative and (soft) related text.

When measured by relevance (i.e., only relevant documents were the target), the top runs used similar information, though all top runs used the (hard) related text.

For passage retrieval evaluation, the best runs were usually automatic (though the second ranked run was manual), used the title and scription, incorporated a clarification form, and did passage retrieval. Interestingly, the fifth ranked run was a document run with no passages marked. Some sites were able to find advantage to the geography and genre metadata, and some used related text and narrative. Note that related text (of both kinds) was more often used in top performing document retrieval systems than in top performing passage retrieval systems.

No top run by any of the measures used the familiarity field.

12 Conclusion

The second year of the HARD track appears to have been much more productive for most sites. With better training data and a clearer task definition earlier, groups were able to carry out more careful and interesting research.

The HARD track will continue in TREC 2005. Funding considerations have forced the removal of passage retrieval from the evaluation. Topics deemed by the Robust track to be difficult will be used rather than developing new topics, though they will be judged against a new corpus. Familiarity metadata will be collected, but not used in any particular way by the annotators.

Acknowledgments

The coordination of the HARD track at the CIIR would not have been possible without the help of Fernando Diaz, Mark Smucker, and Courtney Wade. The way the track was organized was by consensus of participating researchers and they all deserve credit for the shape the track eventually took.

The work at the CIIR was supported in part by the Center for Intelligent Information Retrieval, in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903, and in part by NSF grant number IIS-9907018. Any opinions, findings and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsor.

References

[Abdul-Jaleel et al., 2005] Abdul-Jaleel, N., Allan, J., Croft, W. B., Diaz, F., Larkey, L., Li, X., Smucker, M. D., and Wade, C. (2005). UMass at TREC 2004: Novelty and HARD. In *Proceedings of TREC 2004*. Appears in this volume.

[Belkin et al., 2005] Belkin, N., Chaleva, I., Cole, M., Li, Y.-L., Liu, L., Liu, Y.-H., Muresan, G., Smith, C., Sun, Y., Yuan, X.-J., and Zhang, X.-M. (2005). Rutgers' HARD track experiences at TREC 2004. In *Proceedings of TREC 2004*. Appears in this volume.

[Buckley and Voorhees, 2004] Buckley, C. and Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *Proceedings of SIGIR*, pages 25–32.

[Evans et al., 2005] Evans, D. A., Bennett, J., Montgomery, J., Sheftel, V., Hull, D. A., and Shanahan, J. G. (2005). TREC-2004 HARD-track experiments in clustering. In *Proceedings of TREC 2004*. Appears in this volume.

[Harper et al., 2005] Harper, D. J., Muresan, G., Liu, B., Koychev, I., Wettschereck, D., and Wiratanga, N. (2005). The Robert Gordon University's HARD track experiments at TREC 2004. In *Proceedings of TREC 2004*. Appears in this volume.

[He et al., 2005] He, D., Demner-Fushman, D., Oard, D. W., Karakos, D., and Dhudanpur, S. (2005). Improving passage retrieval using interactive elicitation and statistical modeling. In *Proceedings of TREC 2004*. Appears in this volume.

[Huang et al., 2005] Huang, X., Huang, Y. R., Wen, M., and Zhong, M. (2005). York University at TREC 2004: HARD and genomics tracks. In *Proceedings of TREC 2004*. Appears in this volume.

[Jiang and Zhai, 2005] Jiang, J. and Zhai, C. (2005). UIUC in HARD 2004 – passage retrieval using HMMs. In *Proceedings of TREC 2004*. Appears in this volume.

[Kelly et al., 2005] Kelly, D., Dollu, V. D., and Fu, X. (2005). University of North Carolina's HARD track experiments at TREC 2004. In *Proceedings of TREC 2004*. Appears in this volume.

[Levow, 2005] Levow, G.-A. (2005). University of Chicago at TREC 2004: HARD track. In *Proceedings of TREC 2004*. Appears in this volume.

[Rode and Hiemstra, 2005] Rode, H. and Hiemstra, D. (2005). Conceptual language models for context-aware text retrieval. In *Proceedings of TREC 2004*. Appears in this volume.

[Sun et al., 2005] Sun, L., Zhang, J., and Sun, Y. (2005). ISCAS at TREC-2004: HARD track. In *Proceedings of TREC 2004*. Appears in this volume.

[Vechtomova and Karamuftuoglu, 2005] Vechtomova, O. and Karamuftuoglu, M. (2005). Approaches to high accuracy retrieval: Phrase-based search experiments in the HARD track. In *Proceedings of TREC 2004*. Appears in this volume.

[Voorhees, 2005] Voorhees, E. M. (2005). Overview of the TREC 2004 question answering track. In *Proceedings of TREC 2004*. Appears in this volume.

[Yang et al., 2005] Yang, K., Yu, N., Wead, A., La Rowe, G., Li, Y.-H., Friend, C., and Lee, Y. (2005). WIDIT in TREC-2004 genomics, HARD, robust and web tracks. In *Proceedings of TREC 2004*. Appears in this volume.

[Zaragoza et al., 2005] Zaragoza, H., Craswell, N., Taylor, M., Saria, S., and Robertson, S. (2005). Microsoft Cambridge at TREC-13: Web and HARD tracks. In *Proceedings of TREC 2004*. Appears in this volume.

Overview of the TREC 2004 Novelty Track

Ian Soboroff
National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

TREC 2004 marks the third and final year for the novelty track. The task is as follows: Given a TREC topic and an ordered list of documents, systems must find the relevant and novel sentences that should be returned to the user from this set. This task integrates aspects of passage retrieval and information filtering. As in 2003, there were two categories of topics – events and opinions – and four subtasks which provided systems with varying amounts of relevance or novelty information as training data. This year, the task was made harder by the inclusion of some number of irrelevant documents in document sets. Fourteen groups participated in the track this year.

1 Introduction

The novelty track was introduced in TREC 2002 [1]. The basic task is as follows: given a topic and an ordered set of documents segmented into sentences, return sentences that are both relevant to the topic and novel given what has already been seen. This task models an application where a user is skimming a set of documents, and the system highlights new, on-topic information.

There are two problems that participants must solve in the novelty track. The first is identifying relevant sentences, which is essentially a passage retrieval task. Sentence retrieval differs from document retrieval because there is much less text to work with, and identifying a relevant sentence may involve examining the sentence in the context of those surrounding it. We have specified the unit of retrieval as the sentence in order to standardize the task across a variety of passage retrieval approaches, as well as to simplify the evaluation.

The second problem is that of identifying those relevant sentences that contain new information. The operational definition of "new" is information that has not appeared previously in this topic's set of documents. In other words, we allow the system to assume that the user is most concerned about finding new information in this particular set of documents and is tolerant of reading information he already knows because of his background knowledge. Since each sentence adds to the user's knowledge, and later sentences are to be retrieved only if they contain new information, novelty retrieval resembles a filtering task.

To allow participants to focus on the filtering and passage retrieval aspects separately, the novelty track has four different tasks. The base task was to identify all relevant and novel sentences in the documents. The other tasks provided varying amounts of relevant and novel sentences as training data.

The track has changed slightly from year to year. The first run in 2002 used old topics and relevance judgments, with sentences judged by new assessors [1]. TREC 2003 included separate tasks, made the document ordering chronological rather than relevance-based, and introduced new topics and the different topic types [2]. This year, the major change is the inclusion (or perhaps re-introduction) of irrelevant documents into the document sets.

2 Input Data

The documents for the novelty track are taken from the AQUAINT collection. This collection is unique in that it contains three news sources from overlapping time periods: New York Times News Service (Jun 1998 – Sep 2000), AP (also Jun 1998 – Sep 2000), and Xinhua News Service (Jan 1996 – Sep 2000). As a result, this collection exhibits greater redundancy than other TREC collections, and thus less novel information, increasing the realism of the task.

The NIST assessors created fifty new topics for the 2004 track. As was done last year, the topics were of two types. Twenty-five topics concerned events, such as India and Pakistan's nuclear tests in 1998, and twenty-five topics focused on opinions about controversial subjects such as the safety of irradiated food and the so-called "abortion pill" RU-486. The topic type was indicated in the topic description by a <toptype> tag. The assessors, in creating their topics, searched the AQUAINT collection using WebPRISE, NIST's IR system, and collected 25 documents they deemed to be relevant to the topic. They also labeled some documents as irrelevant, and all documents judged irrelevant and ranked above the 25 relevant documents were included in the document sets. Note that this means that the irrelevant documents are close matches to the relevant ones, and not random irrelevant documents.

Once selected, the documents were ordered chronologically. (Chronological ordering is achieved trivially in the AQUAINT collection by sorting document IDs.) The documents were then split into sentences, each sentence receiving an identifier, and all sentences were concatenated together to produce the document set for a topic.

3 Task Definition

There are four tasks in the novelty track:

Task 1. Given the set of documents for the topic, identify all relevant and novel sentences.

Task 2. Given the relevant sentences in all documents, identify all novel sentences.

Task 3. Given the relevant and novel sentences in the first 5 documents **only**, find the relevant and novel sentences in the remaining documents. Note that since some documents are irrelevant, there *may not be* any relevant or novel sentences in the first 5 documents for some topics.

Task 4. Given the relevant sentences from all documents, and the novel sentences from the first 5 documents, find the novel sentences in the remaining documents.

These four tasks allowed the participants to test their approaches to novelty detection given different levels of training: none, partial, or complete relevance information, and none or partial novelty information.

Participants were provided with the topics, the set of sentence-segmented documents, and the chronological order for those documents. For tasks 2-4, training data in the form of relevant and novel "sentence qrels" were also given. The data were released and results were submitted in stages to limit "leakage" of training data between tasks. Depending on the task, the system was to output the identifiers of sentences which the system determined to contain relevant and/or novel relevant information.

4 Evaluation

4.1 Creation of truth data

Judgments were created by having NIST assessors manually perform the first task. From the concatenated document set, the assessor selected the relevant sentences, then selected those relevant sentences that were novel. Each topic was independently judged by two different assessors, the topic author and a "secondary" assessor, so that the effects of different human opinions could be assessed.

The assessors only judged sentences in the relevant documents. Since, by the definition of relevance in TREC, a document containing any relevant information would itself be relevant, the assessors would not miss any relevant information by not judging the sentences in the irrelevant documents. This does give the second assessor some advantage against systems attempting task 1, since the assessor was not confronted with irrelevant documents in the sentence judging phase.

Since the novelty task requires systems to automatically select the same sentences that were selected manually by the assessors, it is important to analyze the characteristics of the manually-created truth data in order to better understand the system results. The first novelty track topics (in 2002) were created using topics from old TRECs and relevant documents from manual TREC runs, and the sentences judgments were made by new assessors. Those topics had very few relevant sentences and consequently nearly every relevant sentence was novel. Last year's topics, which were each newly developed and judged by a single assessor, resulted in topics with much more reasonable levels of relevant and new information. This year the inclusion of irrelevant documents means that fewer sentences are relevant. Somewhat surprisingly, perhaps, the fraction of relevant sentences which are novel is lower than last year as well.

Table 1 shows the number of relevant and novel sentences selected for each topic by each of the two assessors who worked on that topic. The column marked "assr-1" precedes the results for the primary assessor, whereas "assr-2" precedes those of the secondary assessor. The column marked "rel" is the number of sentences selected as relevant; the next

Table 1: Analysis of relevant and novel sentences by topic

Topic	type	sents	assr-1	rel	% total	new	% rel	assr-2	rel	% total	new	% rel
N51	E	669	C	107	15.99	26	24.30	B	112	16.74	38	33.93
N53	E	667	E	106	15.89	31	29.25	C	136	20.39	86	63.24
N54	E	1229	E	198	16.11	71	35.86	B	384	31.24	224	58.33
N55	E	536	C	56	10.45	21	37.50	E	96	17.91	46	47.92
N56	E	1904	E	196	10.29	103	52.55	A	133	6.99	47	35.34
N57	E	378	B	21	5.56	10	47.62	D	170	44.97	116	68.24
N59	E	855	D	214	25.03	86	40.19	C	152	17.78	62	40.79
N64	E	679	C	214	31.52	140	65.42	A	228	33.58	64	28.07
N68	E	1331	B	200	15.03	45	22.50	E	210	15.78	82	39.05
N69	E	367	D	169	46.05	55	32.54	B	122	33.24	59	48.36
N72	E	1007	B	147	14.60	43	29.25	D	144	14.30	48	33.33
N73	E	380	D	268	70.53	139	51.87	A	164	43.16	93	56.71
N74	E	502	D	240	47.81	107	44.58	C	129	25.70	79	61.24
N79	E	1580	C	199	12.59	69	34.67	D	188	11.90	116	61.70
N80	E	447	E	74	16.55	48	64.86	B	104	23.27	51	49.04
N81	E	684	A	173	25.29	31	17.92	C	236	34.50	167	70.76
N82	E	1152	C	355	30.82	165	46.48	B	100	8.68	44	44.00
N83	E	816	A	250	30.64	62	24.80	E	227	27.82	122	53.74
N85	E	1419	B	181	12.76	95	52.49	E	116	8.17	59	50.86
N87	E	1026	D	476	46.39	163	34.24	C	369	35.96	231	62.60
N88	E	708	C	312	44.07	171	54.81	E	307	43.36	131	42.67
N90	E	1971	B	529	26.84	168	31.76	D	762	38.66	310	40.68
N92	E	879	B	188	21.39	172	91.49	A	199	22.64	83	41.71
N95	E	627	E	78	12.44	36	46.15	D	168	26.79	108	64.29
N98	E	408	C	171	41.91	65	38.01	A	267	65.44	67	25.09
N52	O	1018	B	103	10.12	55	53.40	C	298	29.27	202	67.79
N58	O	1346	A	146	10.85	42	28.77	C	252	18.72	163	64.68
N60	O	948	B	172	18.14	64	37.21	A	257	27.11	79	30.74
N61	O	1150	A	70	6.09	21	30.00	B	78	6.78	40	51.28
N62	O	3132	E	89	2.84	45	50.56	D	97	3.10	79	81.44
N63	O	518	B	49	9.46	21	42.86	E	84	16.22	55	65.48
N65	O	705	B	95	13.48	61	64.21	C	113	16.03	90	79.65
N66	O	795	A	195	24.53	25	12.82	E	286	35.97	137	47.90
N67	O	423	E	113	26.71	72	63.72	C	109	25.77	82	75.23
N70	O	1030	D	94	9.13	31	32.98	E	237	23.01	104	43.88
N71	O	908	B	62	6.83	28	45.16	A	127	13.99	28	22.05
N75	O	2922	B	169	5.78	100	59.17	C	284	9.72	245	86.27
N76	O	1697	A	217	12.79	51	23.50	D	118	6.95	39	33.05
N77	O	1144	D	74	6.47	23	31.08	B	102	8.92	36	35.29
N78	O	1308	A	145	11.09	59	40.69	B	59	4.51	25	42.37
N84	O	1363	D	101	7.41	31	30.69	E	153	11.23	80	52.29
N86	O	493	D	67	13.59	33	49.25	A	96	19.47	46	47.92
N89	O	1271	B	204	16.05	130	63.73	A	181	14.24	61	33.70
N91	O	1473	B	112	7.60	51	45.54	D	123	8.35	99	80.49
N93	O	1017	B	181	17.80	56	30.94	E	255	25.07	129	50.59
N94	O	1099	E	102	9.28	59	57.84	A	91	8.28	46	50.55
N96	O	1328	A	131	9.86	60	45.80	D	61	4.59	45	73.77
N97	O	1416	A	123	8.69	31	25.20	B	122	8.62	89	72.95
N99	O	1192	C	259	21.73	131	50.58	D	495	41.53	341	68.89
N100	O	530	E	148	27.92	52	35.14	B	152	28.68	78	51.32

column, "% total", is the percentage of the total set of sentences for that topic that were selected as relevant. The column marked "new" gives the number of sentences selected as novel; the next column, "% rel", is the percentage of relevant sentences that were marked novel. The column "sents" gives the total number of sentences for that topic, and "type" indicates whether the topic is about an event (**E**) or about opinions on a subject (**O**).

Because this year's document sets include irrelevant documents, the fraction of relevant sentences is less than half that of last year: a mean of 19.2%, compared with 41.1% in TREC 2003. However, the amount of novel information as a fraction of relevant is also lower: a 42% this year vs. 64.6% in TREC 2003. This was somewhat surprising as the collection and topic types are the same, and the topics have the same number of relevant documents. Beyond simple intertopic variation, these topics just have more redundant information.

Opinion topics tended to have fewer relevant sentences than event topics. 25.9% of sentences in event topics were relevant, compared to only 12.6% in opinion topics. Even though the topics are about opinions, the documents are still news stories and thus include current events and background information in addition to the relevant opinion material. The fraction of relevant sentences which were novel was the same for both types, 42%.

In examining assessor effects, this year we were able to achieve much better balance in the second round of assessing, with each assessor judging five topics written by someone else. Overall, the assessors tended to find about the same amount of relevant information whether they were judging their own topics or someone else's (19.2% for their own topics vs. 21.7% in the second round, not significant by a t-test), but identified more novel sentences (42% vs. 52.6%, significant at $p = 0.0009$). We have not made a detailed analysis of how the assessors differed in particular judgments or in their judging patterns.

In summary, the topics for this year seem comparable in quality to the TREC 2003 topics, with minimal assessor effects. The inclusion of irrelevant documents makes the task this year harder for systems, and thus the two topic sets should not be combined.

4.2 Scoring

The sentences selected manually by the NIST assessor who created the topic were considered the truth data. The judgments by the secondary assessor were taken as a human baseline performance in the first task. Relevant and novel sentence retrieval have each been evaluated separately.

Because relevant and novel sentences are returned as an unranked set in the novelty track, we cannot use traditional measures of ranked retrieval effectiveness such as mean average precision. One alternative is to use set-based recall and precision. Let M be the number of matched sentences, i.e., the number of sentences selected by both the assessor and the system, A be the number of sentences selected by the assessor, and S be the number of sentences selected by the system. Then sentence set recall is M/A and precision is M/S.

As the TREC filtering tracks have demonstrated, set-based recall and precision do not average well, especially when the assessor set sizes vary widely across topics. Consider the following example as an illustration of the problems. One topic has hundreds of relevant sentences and the system retrieves 1 relevant sentence. The second topic has 1 relevant sentence and the system retrieves hundreds of sentences. The average for both recall and precision over these two topics is approximately .5 (the scores on the first topic are 1.0 for precision and essentially 0.0 for recall, and the scores for the second topic are the reverse), even though the system did precisely the wrong thing. While most real submissions won't exhibit this extreme behavior, the fact remains that set recall and set precision averaged over a set of topics is not a robust diagnostic indicator of system performance. There is also the problem of how to define precision when the system returns no sentences ($S = 0$). Leaving that topic out of the evaluation for that run would mean that different systems would be evaluated over different numbers of topics, while defining precision in the degenerate case to be either 1 or 0 is extreme. (The average scores given in Appendix A defined precision to be 0 when $S = 0$ since that seems the least evil choice.)

To avoid these problems, the primary measure for novelty track runs is the F measure. The F measure (from van Rijsbergen's E measure) is a function of set recall and precision, together with a parameter β which determines the relative importance of recall and precision. A β value of 1, indicating equal weight, is used in the novelty track. $F_{\beta=1}$ is given as:

$$\mathrm{F} = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}}$$

Alternatively, this can be formulated as

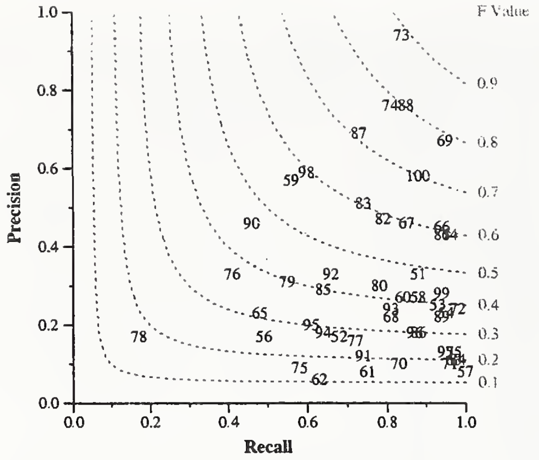


Figure 1: The F measure, plotted according to its precision and recall components. The lines show contours at intervals of 0.1 points of F. The black numbers are per-topic scores for one novelty track run.

$$F = \frac{2 \times (\#\ \text{relevant sentences retrieved})}{(\#\ \text{retrieved sentences}) + (\#\ \text{relevant sentences})}$$

For any choice of β, F lies in the range $[0, 1]$, and the average of the F measure is meaningful even when the judgment sets sizes vary widely. For example, the F measure in the scenario above is essentially 0, an intuitively appropriate score for such behavior. Using the F measure also deals with the problem of what to do when the system returns no sentences since recall is 0 and the F measure is legitimately 0 regardless of what precision is defined to be.

Note, however, that two runs with equal F scores do not indicate equal precision and recall. Figure 1 illustrates the shape of the F measure in precision-recall space. An F score of 0.5, for example, can describe a range of precision and recall scores. Figure 1 also includes the per-topic scores for a particular run are also plotted. It is easy to see that topics 98, 83, 82, and 67 exhibit a wide range of performance, but all have an F score of close to 0.6. Thus, two runs with equal F scores may be performing quite differently, and a difference in F scores can be due to changes in precision, recall, or both.

5 Participants

Table 2 lists the 14 groups that participated in the TREC 2004 novelty track. Nearly every group attempted the first two tasks, but tasks three and four were less popular than last year, with only 8 groups participating in each (compared to 10 last year). The rest of this section contains short summaries submitted by most of the groups about their approaches to the novelty task. For more details, please refer to the group's complete paper in the proceedings.

Most groups took a similar high-level approach to the problem, and the range of approaches is not dramatically different from last year. Relevant sentences were selected by measuring similarity to the topic, and novel sentences by dissimilarity to past sentences. As can be seen from the following descriptions, there is a tremendous variation in how "the topic" and "past sentences" are modeled, how similarity is computed when sentences are involved, and what constitutes the thresholds for relevance and novelty. Many groups tried variations on term expansion to improve sentence similarity, some with more success than others.

5.1 Chinese Academy of Sciences – ICT

In TREC 2004, ICT divided novelty track into four sequential stages. It includes: customized language parsing on original dataset, document retrieval, sentence relevance and novelty detection. In the first preprocessing stage, we applied sentence segmenter, tokenization, part-of-speech tagging, morphological analysis, stop word remover and query analyzer on topics and documents. As for query analysis, we categorized words in topics into description words and content words. Title, description and narrative parts are all merged into query with different weights. In the stage of document and sentence retrieval, we introduced vector space model (VSM) and its variants, probability model (OKAPI) and statistical language model. Based on VSM, we tried various query expansion strategies: pseudo-feedback, term expansion with synset or synonym in WordNet and expansion with highly local co-occurrence terms. With regard to the novelty stage, we defined three types of new degree: word overlapping and its extension, similarity comparison and information gain. In the last three tasks, we used the known results to adjust threshold, estimate the number of results, and turn to classifier, such as inductive and transductive SVM.

5.2 CL Research

The CL Research novelty assessment is based on a full-scale parsing and processing of documents and

Table 2: Organizations participating in the TREC 2004 novelty track

		Runs submitted			
	Run prefix	Task 1	Task 2	Task 3	Task 4
Chinese Academy of Sciences (CAS-ICT)	ICT	5	5	4	5
CL Research	clr	2	1	4	1
Columbia University	nov		5		
Dublin City University	cdvp	5	5		
IDA / Center for Computing Science	ccs	5	5	4	
Institut de Recherche en Informatique de Toulouse	IRIT	5	2	5	
Meiji University	Meiji	3	5	3	5
National Taiwan University	NTU	5	5		
Tsinghua University	THUIR	5	5	5	5
University of Iowa	UIowa	5	5	5	5
University of Massachusetts	CIIR	2	5	3	
University of Michigan	umich	5	5	5	4
Université Paris-Sud / LRI	LRI	5	5		
University of Southern California-ISI	ISI	5			

topic descriptions (titles, descriptions, and narratives) into an XML representation characterizing discourse structure, syntactic structure (particularly noun, verb, and prepositional phrases), and semantic characterizations of open-class words. Componential analysis of the topic narratives was used as the basis for identifying key words and phrases in the document sentences. Several scoring metrics were used to determine the relevance for each sentence. In TREC 2004, the presence of communication nouns and verbs in the narratives was used to expand relevance assessments, by identifying communication verbs in the sentences. This significantly increased recall over TREC 2004, without a significant degradation of precision. CL Research's novelty component was unchanged, but precision on Task 2 was considerably lower. This lower precision was observed in other tasks as well, and perhaps reflects the significantly lower scores among all participants. CL Research has set up an evaluation framework to examine the reasons for these lower scores.

5.3 Columbia University

Our system for the novelty track at TREC 2004, SumSeg, for Summary Segmentation, is based on our observations of data we collected for the development of our system to prepare update summaries, or bulletins. We see that new information often appears in text spans of two or more sentences, and at other times, a piece of new information is embedded within a sentence mostly containing previously seen material. In order to capture both types of cases, we avoided direct sentence similarity measures, and took evidence of unseen words as evidence of novelty. We employed a hill climbing algorithm to learn thresholds for how many new words would trigger a novel classification. We also sought to learn different weights for different types of nouns, for example, persons, or locations or common nouns. In addition, we included a mechanism to allow sentences that had few strong content words to "continue" the classification of the previous sentence. Finally, we used two statistics, derived from analysis of the full AQUAINT corpus, to eliminate low-content words. We submitted a total of five runs: two used learned parameters to aim at high precision output, and one aimed at higher recall. Another run was a straightforward vector-space model used as a baseline, and the last was a combination of the high recall run with the vector-space model. Training was done on the 2003 TREC novelty data.

5.4 Dublin City University

This is the first year that DCU has participated in the novelty track. We built three models; the first focused on retrieving the twenty-five documents that were relevant to each topic; the second focused on retrieving relevant sentences from this list of retrieved documents to satisfy each individual topic; the third focused on the detection of novel sentences from this relevant list. In Task1 we used an information retrieval system developed by the CDVP for the terabyte track as a basis for our experiments. This system used the BM25 ranking algorithm. We used various query and document expansion techniques to enhance the performance for sentence level retrieval. In Task 2 we developed two formulas, the ImportanceValue and The NewSentenceValue, which exploit term characteristics using traditional document similarity methods.

5.5 Institut de Recherche en Informatique de Toulouse (IRIT)

In TREC 2004, IRIT modified important features of the strategy that was developed for TREC 2003. These features include both some parameter values, topic expansion and taking into account the order of sentences. According to our method, a sentence is considered relevant if it matches the topic with a certain level of coverage. This coverage depends on the category of the terms used in the texts. Four types of terms have been defined — highly relevant, scarcely relevant, non-relevant (like stop words), highly non-relevant terms (negative terms). Term categorization is based on topic analysis: highly non-relevant terms are extracted from the narrative parts that describe what will be a non-relevant document. The three other types of terms are extracted from the rest of the query and are distinguished according to the score they obtain. The score is based both on the term occurrence and on the topic part they belong to (Title, descriptive, narrative). Additionally we increase the score of a sentence when the previous sentence is relevant. When topic expansion is applied, terms from relevant sentences (task 3) or from the first retrieved sentences (task 1) are added to the initial terms. With regard to the novelty part, a sentence is considered as novel if its similarity with each of the previously processed and *selected as novel* sentences does not exceed a certain threshold. In addition, this sentence should not be too similar to a virtual sentence made of the n best-matching sentences.

5.6 University of Iowa

Our system for novelty this year comprises three distinct variations. The first is a refinement of that used for last year involving named entity occurrences and functions as a comparative baseline. The second variation extends the baseline system in an exploration of the connection between word sense and novelty through two alternatives. The first alternative attempts to address the semantics of novelty by expanding all noun phrases (and contained nouns) to their corresponding WordNet synset IDs, and subsequently using synset IDs for novelty comparisons. The second alternative performs word sense disambiguation using an ensemble scheme to establish whether the additional computational overhead is warranted by an increase in performance over simple sense expansion.

The third variation involves more 'traditional' similarity schemes in the positive sense for relevance and the negative sense for novelty. SMART is first used to identify the top 25 documents and then judges relevance at the sentence level to generate a preliminary pool of candidates and then incrementally extends a matched terminology vector. The matched term vector is then used to rematch candidate sentences. Only similarities below a threshold - and hence possessing sufficient dissimilarity are declared novel.

5.7 University of Massachusetts

For relevant sentences retrieval, our system treated sentences as documents and took the words in the title field of the topics as queries. TFIDF techniques with selective feedback were used for retrieving relevant sentences. Selective pseudo feedback means pseudo feedback was performed on some queries but not on other queries based on an automatic analysis on query words across different topics. Basically, a query with more focused query words that rarely appear in relevant documents related to other queries was likely to have a better performance without pseudo feedback. Selective relevance feedback was performed when relevance judgment of top five documents was available as for Task 3. Whether to performance relevance feedback on a query was determined by the comparison between the performance with and without relevance feedback in the top five documents for this query.

For identifying novel sentences, our system started with the sentences returned from the relevant sentences retrieval. The cosine similarity between a sentence and each previous sentence was calculated. The maximum similarity was used to eliminate redundant sentences. Sentences with a maximum similarity greater than a preset threshold were treated as redundant sentences. The value of the same threshold for all topics was tuned with the TREC 2003 track data when no judgment was available. The value of the threshold for each topic was trained with the training data when given the judgment of the top five documents. In addition to the maximum similarity, new words and named entities were also considered in identifying novel sentences.

5.8 University of Michigan

We view a cluster of documents as a graph, where each node is a sentence. We define an edge between two nodes where the cosine similarity between the corresponding two sentences is above a predefined threshold. After this graph is constructed, we find the eigenvector centrality score for each sentence by using a power method, which also corresponds to the stationary distribution of the stochastic graph.

To find the relevant sentences, we compute eigenvector centrality for each sentence together with some other heuristic features such as the similarity between the sentence and the title and/or description of the topic. To find the new sentences, we form the cosine similarity graph that consists of only the relevant sentences. Since the order of the sentences is important, unlike the case in finding the relevant sentences, we form a directed graph where every sentence can only point to the sentences that come after and are similar to it. The more incoming edges a sentence has, the more repeated information it contains. Therefore, the sentences with low centrality scores are considered as new. The system is trained on 2003 data using maximum entropy or decision lists.

5.9 Université Paris-Sud – LRI

The text-mining system we are building deals with the specific problem of identifying the instances of relevant concepts found in the texts. This has several consequences. We develop a chain of linguistic treatment such that the n-th module improves the semantic tagging of the $(n-1)$-th. This chain has to be friendly toward at least two kinds of experts: a linguistic expert, especially for the modules dealing mostly with linguistic problems (such as correcting wrong grammatical tagging), and a field expert for the modules dealing mostly with the meaning of group of words. Our definition of friendliness includes also developing learning procedures adapted to various steps of the linguistic treatment, mainly for grammatical tagging, terminology, and concept learning. In our view, concept learning requires a special learning procedure that we call Extensional Induction. Our interaction with the expert differs from classical supervised learning, in that the expert is not simply a resource who is only able to provide examples, and unable to provide the formalized knowledge underlying these examples. This is why we are developing specific programming languages which enable the field expert to intervene directly in some of the linguistic tasks. Our approach is thus not particularly well adapted to the TREC competition, but our results show that the whole system is functional and that it provides usable information.

In this TREC competition we worked at two levels of our complete chain. In one level, we stopped the linguistic treatment at the level of terminology (i.e., detecting the collocations relevant to the text). Relevance was then defined as the appearance of the same terms in the task definition (exactly as given by the TREC competition team) and in the texts. Our relatively poor results show that we should have been using relevance definitions extended by human-provided comments. Novelty was defined by a TF*IDF measurement which seems to work quite correctly, but that could be improved by using the expert-defined concepts as we shall now see. The second level stopped the linguistic treatment after the definition of the concepts. Relevance was then defined as the presence of a relevant concept and novelty as the presence of a new concept. For each of the 5 runs, this approach proved to be less efficient than the simpler first one. We noticed however that the use of concepts enabled us to obtain excellent results on specific topics (and extremely bad ones as well) in different runs. We explain these very irregular results by our own lack of ability to define properly the relevant concepts for all the 50 topics since we got our best results on topics that either we understood well (e.g., Pinochet, topic N51) or that were found interesting (e.g., Lt-Col Collins, topic N85).

5.10 University of Southern California – ISI

Our system's two modules recognize relevant event and opinion sentences respectively. We focused mainly on recognizing relevant opinion sentences using various opinion-bearing word lists. This year, each topic contained 25 relevant documents, possibly mixed with additional irrelevant documents. Thus, before proceeding to the next phase we had to separate relevant documents from irrelevant documents. We treat this problem as a standard Information Retrieval (IR) procedure. We used a probabilistic Bayesian inference network model to identify the relevant documents. For opinion topics, we used unigrams as subjectivity clues and built four different systems to generate opinion-bearing word lists. After building these unigram lists, we checked each sentence in the relevant documents for the presence of opinion-bearing words. For event topics, we treat event identification as a traditional document IR task. For the IR part we treat each sentence independently of other sentences and index them accordingly. We thus reduce the problem of event identification to that of sentence retrieval. We choose the description <desc> field for formulating the query.

5.11 Tsinghua University

- Text feature selection and reduction, including using Named Entities, POS-tagging information, and PCA transformation which has been shown to be more effective;
- Improve sentence classification to find relevant information using SVM;
- Efficient sentence redundancy computing, including selected pool approach, tightness restriction factor, and PCA-based cosine similarity measurement;
- Effective result filtering, combining sentence and document similarities.

Several approaches are investigated for the step two of novelty (redundancy reduction): Combining the pool method and sentence to sentence overlap, we have a selected pool method, where unlike in the pool method, not all previously seen sentences are included into the pool, only those thought to be related are included. Tightness restriction to overcome one disadvantage of overlap methods is studied. We observed not all sentences with an overlap of 1 (complete term overlap) are really redundant, so we came up with the idea of tightness restriction which tries to recover highly overlapping but in fact novel sentences. In this method, the ratio of the range of common terms in the previous sentence over the range in the later sentence is used as a statistic. Cosine similarity between sentences after PCA is also investigated, and is proved to be most effective.

6 Results

Figures 2, 4, 5, and 6 show the average F scores for tasks 1, 2, 3, and 4 respectively. For task 1, the system scores are shown alongside the "score" of the secondary assessor, who essentially performed this task (with the caveat that they did not judge sentences in the irrelevant documents). Within the margin of error of human disagreement, the assessor lines can be thought of as representing the best possible performance, and are fairly close to the scores for the second assessor last year.

Last year, the top systems were performing at the level of the second assessor, but this year there is a large gap between the second assessor and the systems. Moreover, nearly all systems had low average precision and high average recall. These two observations seem to imply that systems are much too lenient with what they accept as relevant or novel. Some runs with the lowest F scores actually achieved the highest precision of any run in task 1.

We cannot simply say that the difference in performance is due to the inclusion of irrelevant documents. In task 2, where systems are given all relevant sentences and therefore no interference from irrelevant documents, performance is much lower than in the same task last year. It may be that the systems have overly tuned to the 2003 data.

The systems all scored within a very small range, mostly between $0.36 - 0.4$ for relevant sentences and $0.18 - 0.21$ for novel. Precision is very uniform, but recall varies a lot. Last year, the best runs were also very close to one another; this year, the bottom systems have caught up, but the top systems have not improved very much.

Event topics proved to be easier than opinion topics. Figure 3 illustrates this for task 1, where every run did better on event topics than on opinions. The gap between opinions and events in task 1 is also larger than last year. The same gap exists in task 3, but in tasks 2 and 4, where all relevant sentences are provided, performance on opinion topics is much improved, and some runs do better on opinion topics than events. Thus, we can conclude that identifying sentences containing an opinion remains a hard problem.

Scores for task 2 (Figure 4) and task 4 (Figure 6) are shown against a baseline of returning all relevant sentences as novel. Most systems are doing better than this simplistic approach, both by F score and precision, indicating that the algorithms are successfully being somewhat selective.

It is also surprising how little the systems seem to benefit from training data. Overall scores did not improve between tasks 1 and 3, and from task 2 to task 4, novel sentence retrieval actually decreased significantly (see Figure 7). To be fair, this analysis needs to be balanced across groups, as tasks 3 and 4 had fewer runs and fewer groups participating, and some groups use radically different approaches in the presence of training data. But whereas last year additional training data helped relevant sentence retrieval markedly, this year there is no improvement.

7 Conclusion

This is the third and final year for the novelty track. We have examined a particular kind of novelty detection, that is, finding novel information within documents that the user is reading. This is by no means the only kind of novelty detection. Another important kind is detecting new *events*, which has been studied in the TDT evaluations. There, the user is monitoring a news stream and wants to know when something new, such as a plane crash, is first reported. Yet a third is the problem of returning new stories about a known topic, studied in the TREC filtering track and also in TDT topic tracking and story link detection.

We have seen here that filtering and learning approaches can be applied to detecting novel relevant information within documents, but that it remains a hard problem. Because the unit of interest is a sentence, there is not a lot of data in each unit on which to base the decision. Allowing arbitrary passages would make for a much more complicated evaluation.

The exploration into event and opinion topics has been an interesting and fruitful one. The opinions topics are quite different in this regard than other TREC topics. By mixing the two topic types within each task, we have seen that identifying opinions is hard, even with training data, while detecting new opinions (given relevance) seems analogous to detecting new information about an event.

One interesting footnote to the novelty track has been the use of the data outside the track. We know of two scenarios, namely summarization evaluation in DUC and an opinion detection pilot in AQUAINT, which have made use of topics from the novelty track. It's rewarding to see that this data is proving useful beyond the original scope of the track.

References

[1] Donna Harman. Overview of the TREC 2002 novelty track. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, NIST Special Publication 500-251, pages 46–55, Gaithersburg, MD, November 2002.

[2] Ian Soboroff and Donna Harman. Overview of the TREC 2003 novelty track. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, NIST Special Publication 500-255, Gaithersburg, MD, November 2003.

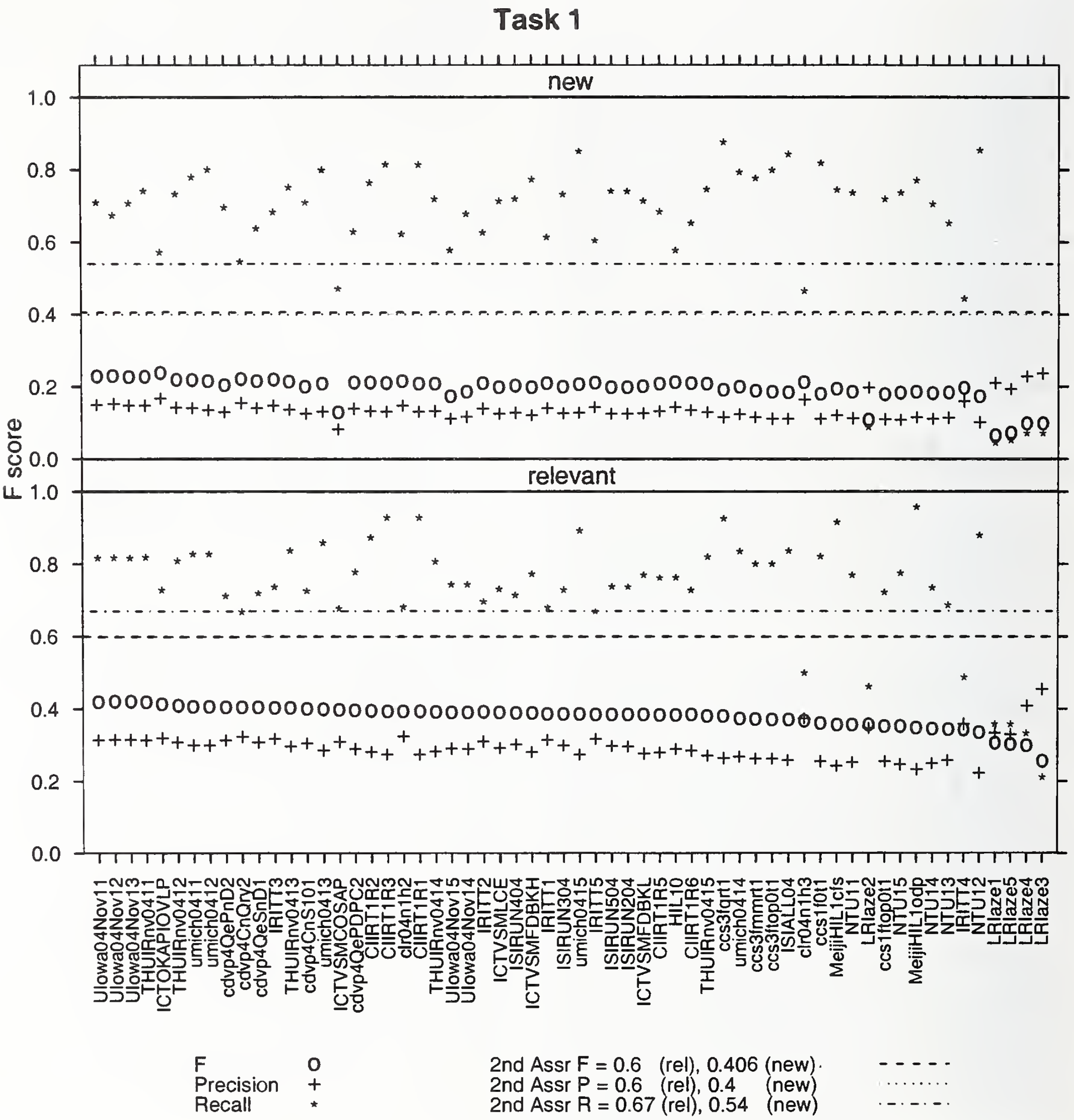


Figure 2: F, precision, and recall scores for Task 1, along with the "average score" of the secondary assessor. Runs are ordered by average F score for relevant sentence retrieval.

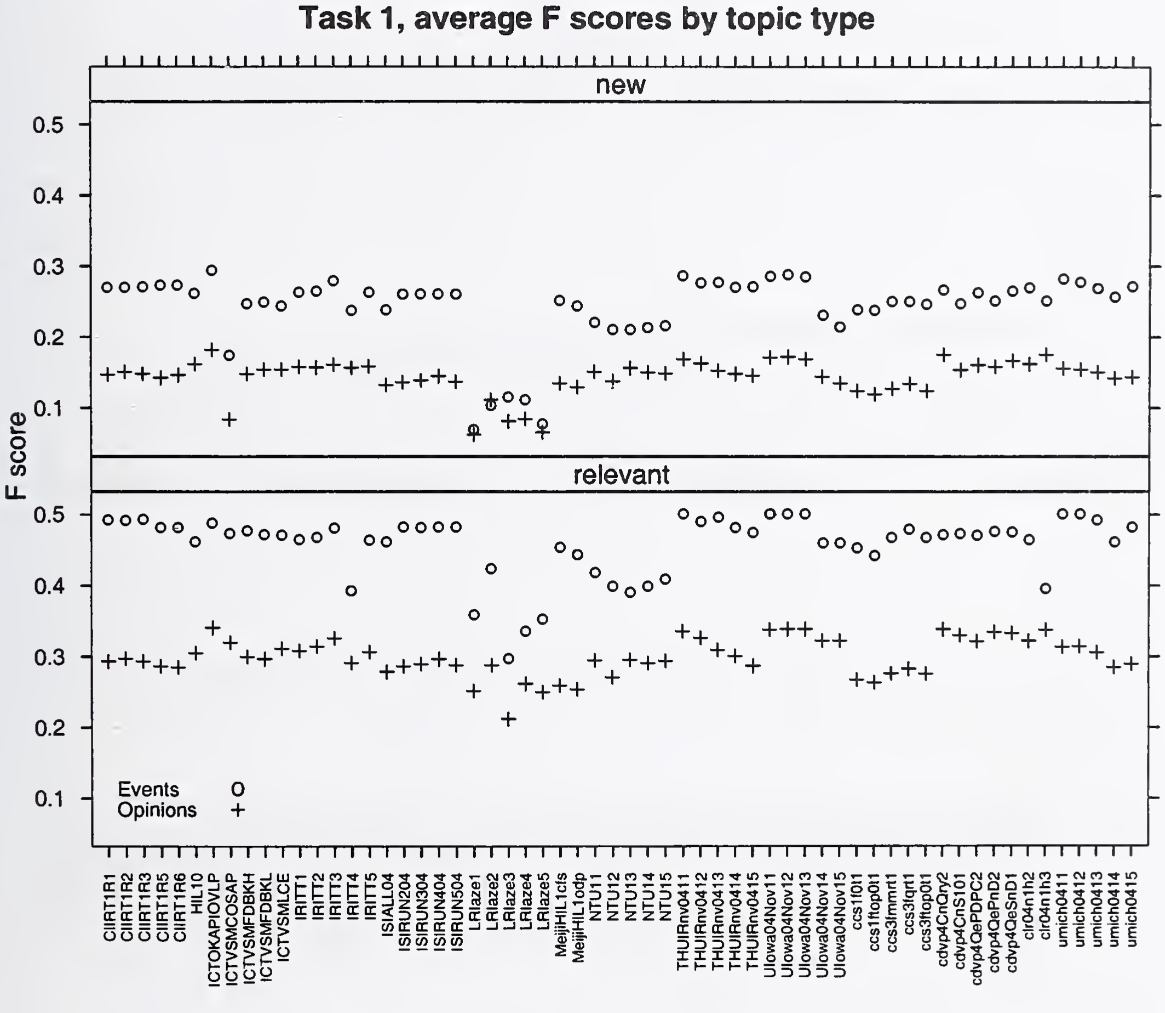


Figure 3: Average F scores per run for opinion and event topic types. Runs are grouped by tag for easier identification.

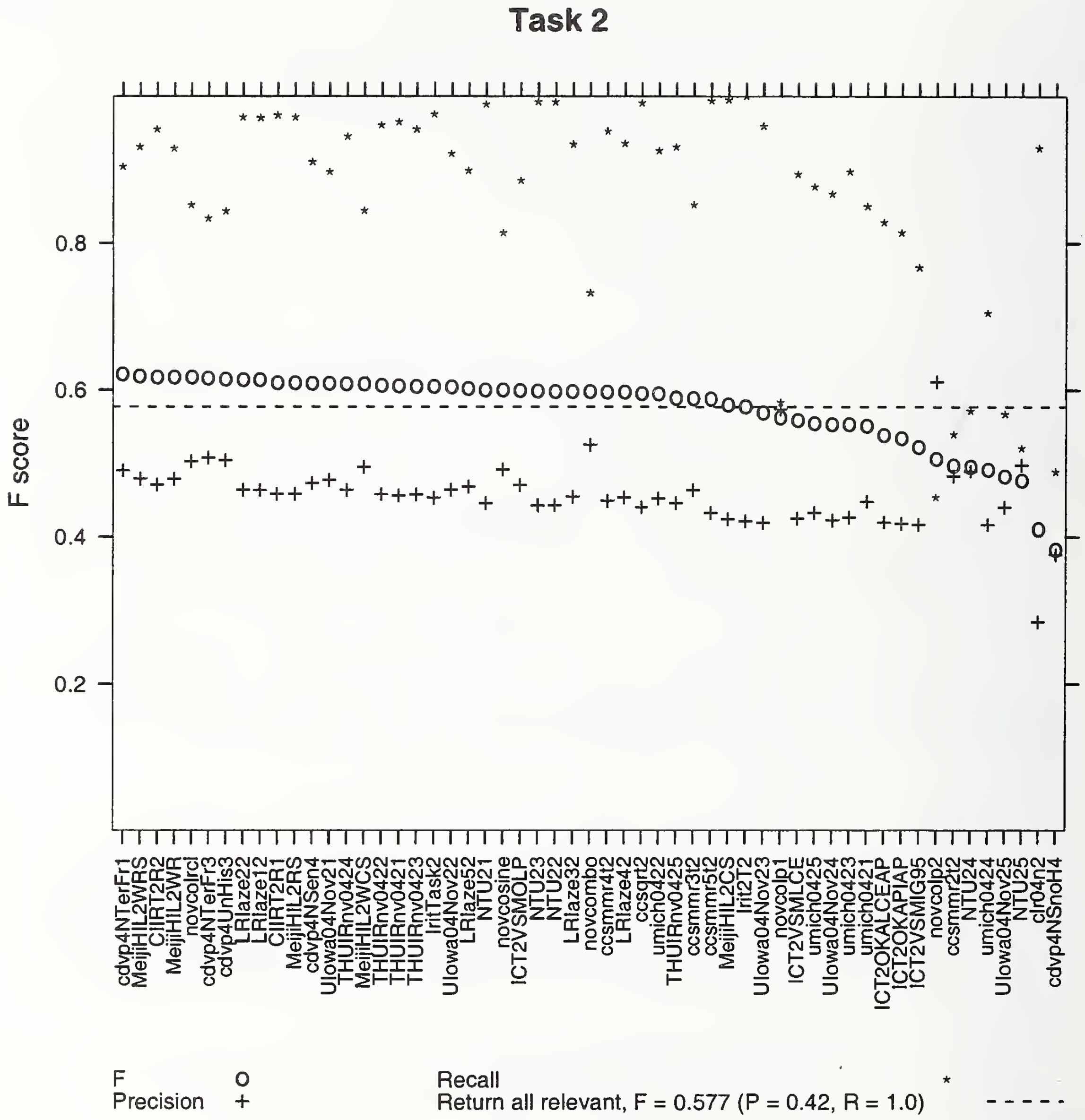


Figure 4: Scores for Task 2, against a baseline of returning all relevant sentences as novel.

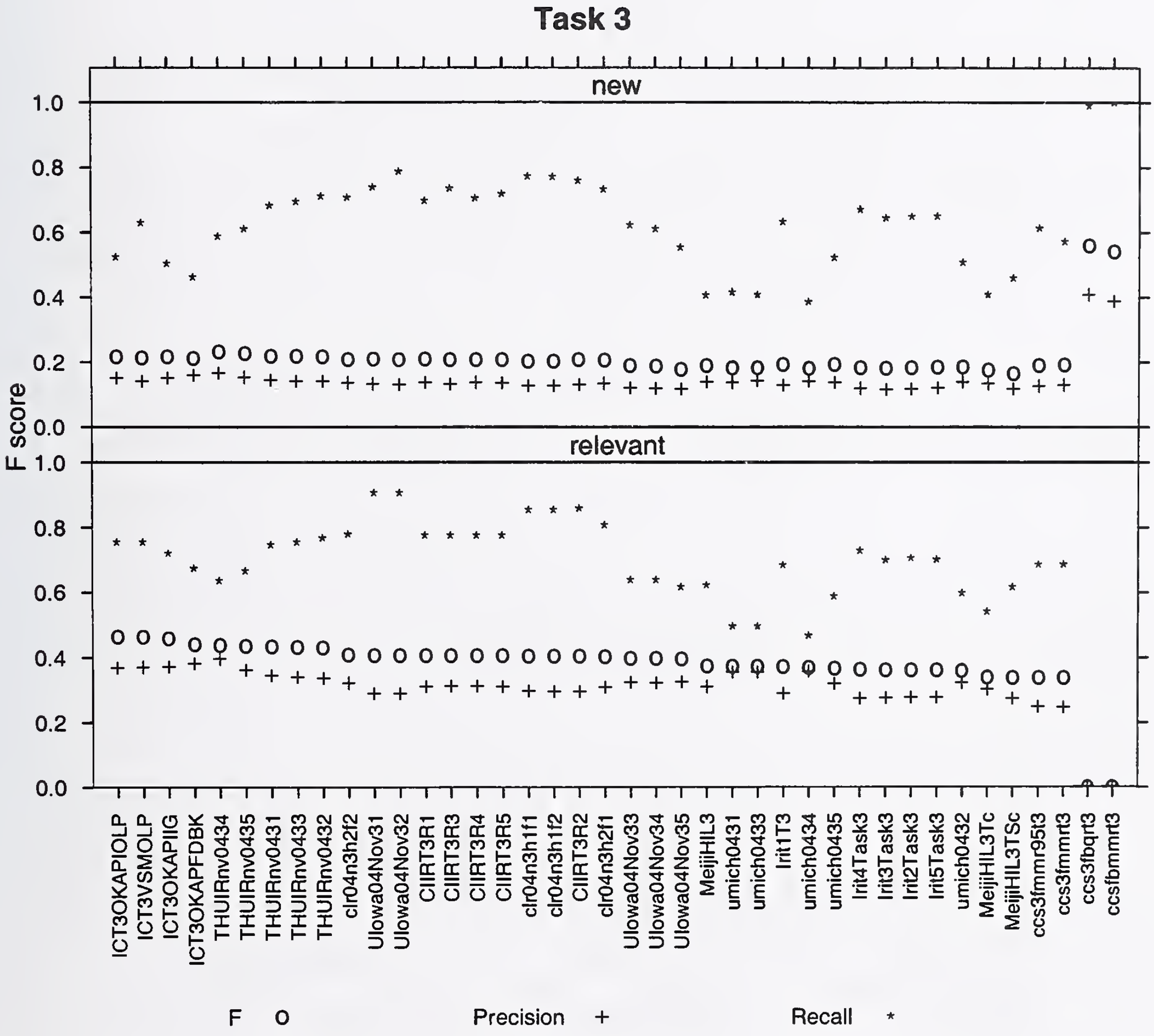


Figure 5: Scores for Task 3, ordered by average F score for relevant sentence retrieval.

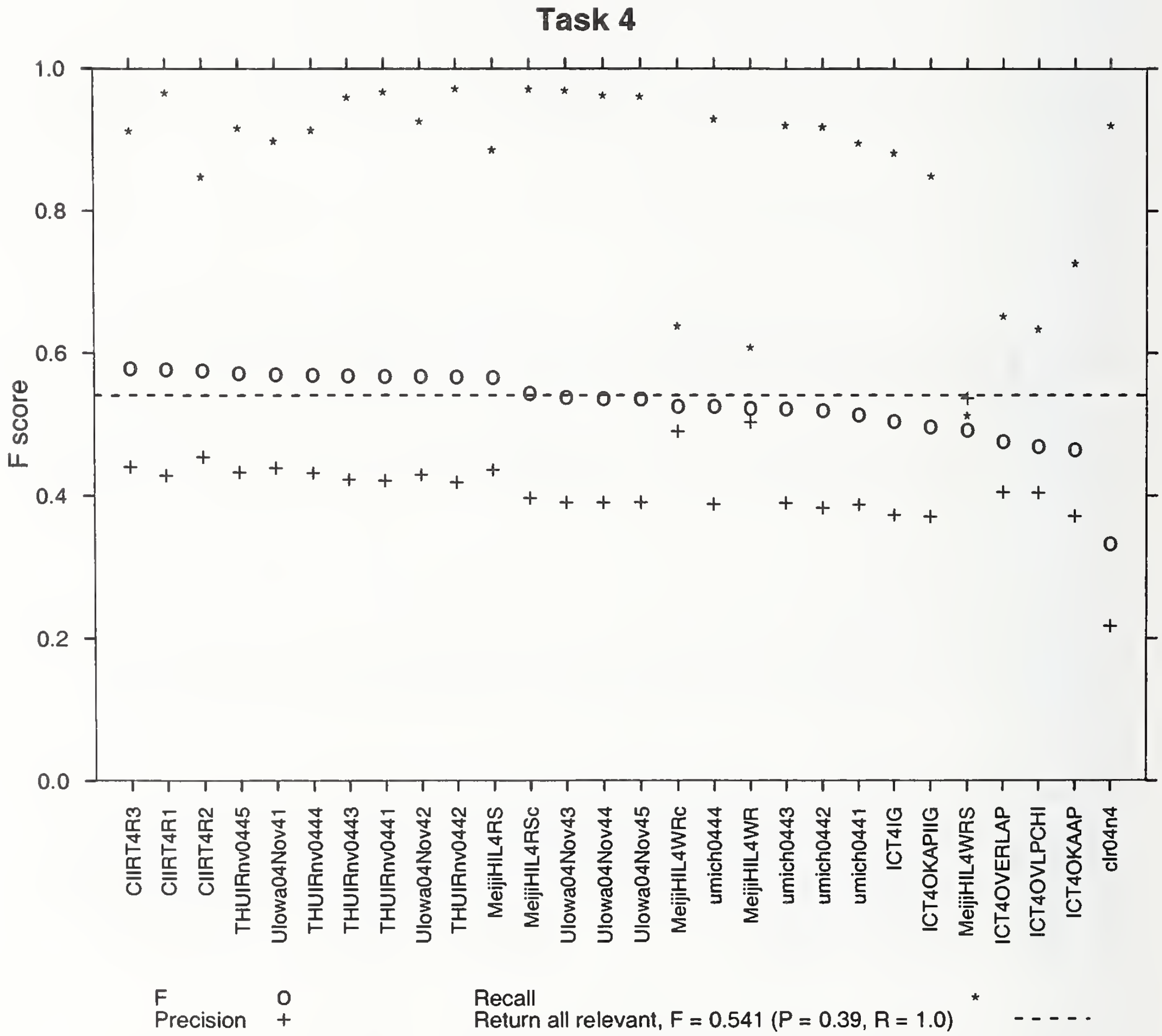


Figure 6: Scores for Task 4, against a baseline of returning all relevant sentences as novel.

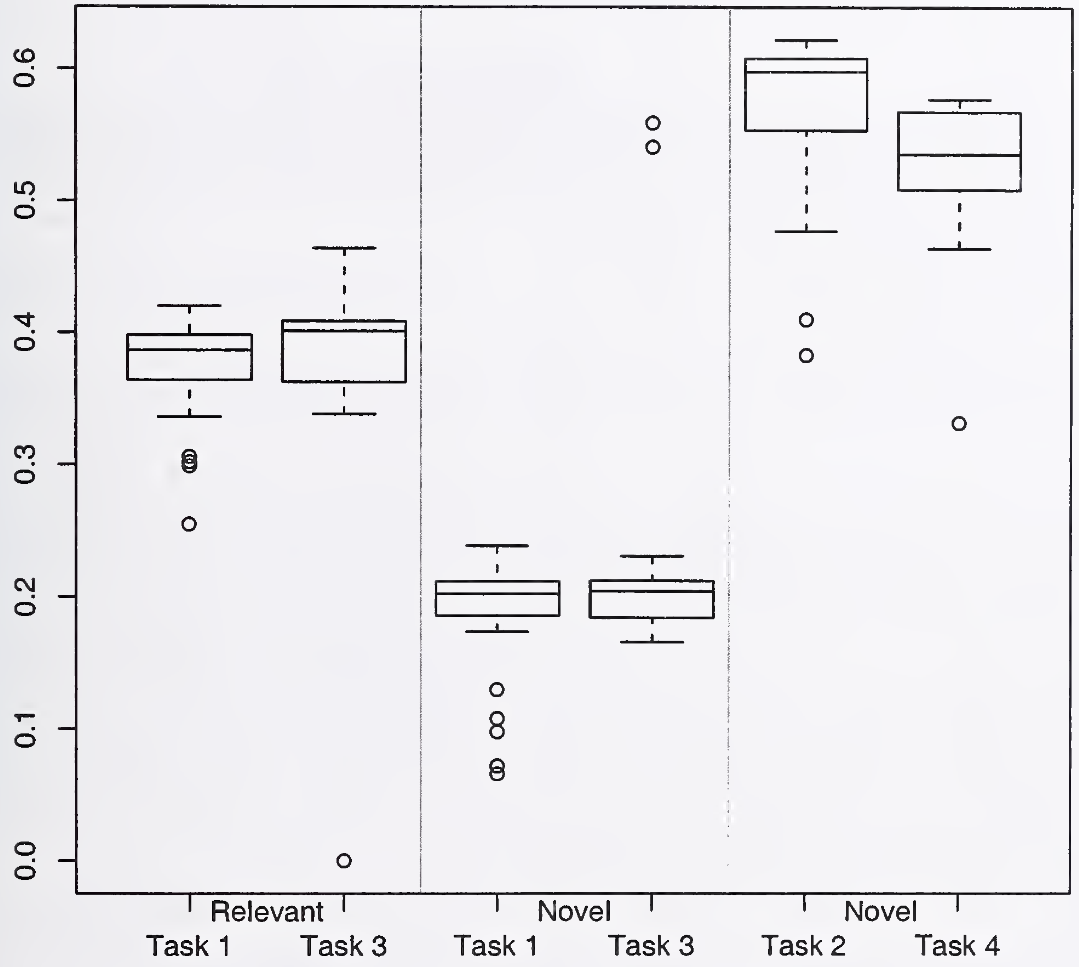


Figure 7: Comparison of F scores between Tasks 1 and 3, and between Tasks 2 and 4.

Overview of the TREC 2004 Question Answering Track

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

The TREC 2004 Question Answering track contained a single task in which question series were used to define a set of targets. Each series contained factoid and list questions and related to a single target. The final question in the series was an "Other" question that asked for additional information about the target that was not covered by previous questions in the series. Each question type was evaluated separately with the final score a weighted average of the different component scores. Applying the combined measure on a per-series basis produces a QA task evaluation that more closely mimics classic document retrieval evaluation.

The goal of the TREC question answering (QA) track is to foster research on systems that return answers themselves, rather than documents containing answers, in response to a question. The track started in TREC-8 (1999), with the first several editions of the track focused on *factoid* questions. A factoid question is a fact-based, short answer question such as *How many calories are there in a Big Mac?*. The task in the TREC 2003 QA track was a combined task that contained list and definition questions in addition to factoid questions [3]. A list question asks for different instances of a particular kind of information to be returned, such as *List the names of chewing gums*. Answering such questions requires a system to assemble an answer from information located in multiple documents. A definition question asks for interesting information about a particular person or thing such as *Who is Vlad the Impaler?* or *What is a golden parachute?*. Definition questions also require systems to locate information in multiple documents, but in this case the information of interest is much less crisply delineated.

The TREC 2003 track was the first large-scale evaluation of list and definition questions, and the results of the track demonstrated that not only are list and definition questions challenging tasks for systems, but they present evaluation challenges as well. Definition task scores contained a relatively large error term in comparison to the size of the difference between scores of different systems. For example, the analysis of the TREC 2003 definition evaluation performed as part of TREC 2003 showed that an absolute difference in scores of 0.1 was needed to have 95% confidence that the comparison represented a true difference in scores when the test set contained 50 questions. Yet relatively few of the runs submitted to TREC 2003 differed by this amount. Reducing the error term requires more definition questions in the test set. The task for the TREC 2004 QA track was designed to accommodate more definition questions while keeping a mix of different question types.

The TREC 2004 test set contained factoid and list questions grouped into different series, where each series had the target of a definition associated with it. Each question in a series asked for some information about the target. In addition, the final question in each series was an explicit "other" question, which was to be interpreted as "Tell me other interesting things about this target I don't know enough to ask directly". This last question is roughly equivalent to the definition questions in the TREC 2003 task.

The reorganization of the combined task into question series has an important additional benefit. Each series is a (limited) abstraction of an information dialog in which the user is trying to define the target. The target and earlier questions in a series provide the context for the current question. Context processing is an important element for question answering systems to possess, but its use has not yet been successfully incorporated into the TREC QA track [2].

The remainder of this paper describes the TREC 2004 QA track in more detail. The next section describes the question series that formed the basis of the evaluation. The following section describes the way the individual question types were evaluated and gives the scores for the runs for that component. Section 3 summarizes the technical

3	Hale Bopp comet		
	3.1	FACTOID	When was the comet discovered?
	3.2	FACTOID	How often does it approach the earth?
	3.3	LIST	In what countries was the comet visible on its last return?
	3.4	OTHER	
21	Club Med		
	21.1	FACTOID	How many Club Med vacation spots are there worldwide?
	21.2	LIST	List the spots in the United States.
	21.3	FACTOID	Where is an adults-only Club Med?
	21.4	OTHER	
22	Franz Kafka		
	22.1	FACTOID	Where was Franz Kafka born?
	22.2	FACTOID	When was he born?
	22.3	FACTOID	What is his ethnic background?
	22.4	LIST	What books did he author?
	22.5	OTHER	

Figure 1: Sample question series from the test set. Series 3 has a THING as a target, series 21 has an ORGANIZATION as a target, and series 22 has a PERSON as a target.

approaches used by the systems to answer the questions. Section 4 looks at the advantages of evaluating runs using a per-series combined score rather than an overall combined score. The final section looks at the future of the track.

1 Question Series

The TREC 2004 QA track consisted of a single task, providing answers for each question in a set of question series. A question series consisted of several factoid questions, zero to two list questions, and exactly one Other question. Associated with each series was a definition target. The series a question belonged to, the order of the question in the series, and the type of each question (factoid, list, or Other) were all explicitly encoded in the XML format used to describe the test set. Example series (minus the XML tags) are shown in figure 1.

The question series were developed as follows. NIST staff searched search engines logs[1] for definition targets. A target was a person, an organization, or thing that was a plausible match for the scenario assumed for the task. The task scenario was the same as in the 2003 track: the questioner was an adult, a native speaker of English, and an "average" reader of US newspapers who was looking for more information about a term encountered while reading the paper.

The set of candidate targets were then given to the assessors, the humans who act as surrogate users and judge the system responses. An assessor selected a target and wrote down questions regarding things he or shee would want to know about the target. The assessor then searched the document collection looking for answers to those questions, plus recording other information about the target that had not asked about but they found interesting. For the most part, the assessors created the questions before doing any searching. However, if the assessor did not know anything about the target (and therefore could create no questions), they first did a Google search to learn about the target, then created questions, and finally searched the document collection. The document collection was the same document set used by the participants as the source of answers, the AQUAINT Corpus of English News Text (LDC catalog number LDC2002T31).

NIST staff reviewed the information found by the assessors and constructed the final question series. Because most questions in the final test set had to contain answers in the document collection, and there needed to be sufficient "other" information for the final question in the series, the final series were heavily edited versions of the assessors' original series. This process proved to be more time-consuming than expected, so a few of the question series were constructed directly from searches of the document collection (i.e., the target was not selected from the logs and the questions were developed only after the search).

[1] The search engine logs were donated by Abdur Chowdhury of AOL and Susan Dumais of Microsoft Research for the TREC 2003 track.

Table 1: Targets of the 65 question series.

S1	Crips	S2	Fred Durst	S3	Hale Bopp comet
S4	James Dean	S5	AARP	S6	Rhodes scholars
S7	agouti	S8	Black Panthers	S9	Insane Clown Posse
S10	prions	S11	the band Nirvana	S12	Rohm and Haas
S13	Jar Jar Binks	S14	Horus	S15	Rat Pack
S16	cataract	S17	International Criminal Court	S18	boxer Floyd Patterson
S19	Kibbutz	S20	Concorde	S21	Club Med
S22	Franz Kafka	S23	Gordon Gekko	S24	architect Frank Gehry
S25	Harlem Globe Trotters	S26	Ice-T	S27	Jennifer Capriati
S28	Abercrombie and Fitch	S29	'Tale of Genji'	S30	minstrel Al Jolson
S31	Jean Harlow	S32	Wicca	S33	Florence Nightingale
S34	Amtrak	S35	Jack Welch	S36	Khmer Rouge
S37	Wiggles	S38	quarks	S39	The Clash
S40	Chester Nimitz	S41	Teapot Dome scandal	S42	USS Constitution
S43	Nobel prize	S44	Sacajawea	S45	International Finance Corporation
S46	Heaven's Gate	S47	Bashar Assad	S48	Abu Nidal
S49	Carlos the Jackal	S50	Cassini space probe	S51	Kurds
S52	Burger King	S53	Conde Nast	S54	Eileen Marie Collins
S55	Walter Mosley	S56	Good Friday Agreement	S57	Liberty Bell 7
S58	philanthropist Alberto Vilar	S59	Public Citizen	S60	senator Jim Inhofe
S61	Muslim Brotherhood	S62	Berkman Center for Internet and Society	S63	boll weevil
S64	Johnny Appleseed	S65	space shuttles		

The final test set contained 65 series; the targets of these series are given in table 1. Of the 65 targets, 23 are PERSONs, 25 are ORGANIZATIONs, and 17 are THINGs. The series contain a total of 230 factoid questions, 56 list questions, and 65 (one per target) Other questions. Each series contains at least 4 questions (counting the Other question), with most series containing 5 or 6 questions. The maximum number of questions in a series is 10.

The question series used in the TREC 2004 track are similar to the QACIAD challenge (Question Answering Challenge for Information Access Dialogue) of NTCIR4 [1]. However, there are some important differences. The heavy editing of the assessors' original questions required to make a usable evaluation test set means the TREC series are not true samples of the assessors' original interests in the target. There were many questions that were eliminated because they did not have answers in the document collection or because they did not meet some other evaluation criterion (for example, the answers for many of the original list questions were not named entities). The TREC series are also not true samples of naturally occurring user-system dialog. In a true dialog, the user would most likely mention answers of previous questions in later questions, but the TREC test set specifically did not do this. This appears as a stilted conversational style when viewed from the perspective of true dialog.

Participants were required to submit retrieval results within one week of receiving the test set. All processing of the questions was required to be strictly automatic. Systems were required to process series independently from one another, and required to process an individual series in question order. That is, systems were allowed to use questions and answers from earlier questions in a series to answer later questions in that same series, but could not "look ahead" and use later questions to help answer earlier questions. As a convenience for the track, NIST made available document rankings of the top 1000 documents per target as produced using the PRISE document retrieval system and the target as the query. Sixty-three runs from 28 participants were submitted to the track.

2 Component Evaluations

The questions in the series were tagged as to which type of question they were because each question type had its own response format and evaluation method. The final score for a run was computed as a weighted average of the component scores. The individual component evaluations for 2004 were identical to those used in the TREC 2003 QA track, and are briefly summarized in this section.

2.1 Factoid questions

The system response for a factoid question was either exactly one [*doc-id*, *answer-string*] pair or the literal string 'NIL'. Since there was no guarantee that a factoid question had an answer in the document collection, NIL was returned by the system when it believed there was no answer. Otherwise, *answer-string* was a string containing precisely an answer to the question, and *doc-id* was the id of a document in the collection that supported *answer-string* as an answer.

Each response was independently judged by two human assessors. When the two assessors disagreed in their judgments, a third adjudicator (a NIST staff member) made the final determination. Each response was assigned exactly one of the following four judgments:

incorrect: the answer string does not contain a right answer or the answer is not responsive;

not supported: the answer string contains a right answer but the document returned does not support that answer;

not exact: the answer string contains a right answer and the document supports that answer, but the string contains more than just the answer or is missing bits of the answer;

correct: the answer string consists of exactly the right answer and that answer is supported by the document returned.

To be responsive, an answer string was required to contain appropriate units and to refer to the correct "famous" entity (e.g., the Taj Mahal casino is not responsive when the question asks about "the Taj Mahal"). NIL responses are correct only if there is no known answer to the question in the collection and are incorrect otherwise. NIL is correct for 22 of the 230 factoid questions in the test set.

The main evaluation score for the factoid component is *accuracy*, the fraction of questions judged correct. Also reported are the recall and precision of recognizing when no answer exists in the document collection. NIL precision is the ratio of the number of times NIL was returned and correct to the number of times it was returned, whereas NIL recall is the ratio of the number of times NIL was returned and correct to the number of times it was correct (22). If NIL was never returned, NIL precision is undefined and NIL recall is 0.0.

Table 2 gives evaluation results for the factoid component. The table shows the most accurate run for the factoid component for each of the top 10 groups. The table gives the accuracy score over the entire set of factoid questions as well as NIL precision and recall scores. In addition, the table reports accuracy for two subsets of the factoid questions: those factoid questions that were the first question in their series (Initial), and those factoid questions that were not the first questions in their series (Non-Initial). As suggested by QACIAD [1], these last two accuracy scores may indicate whether systems had difficulty with context processing in that the first question in a series is usually more fully specified than later questions in a series. (But note there are only 62 initial factoid questions and 168 non-initial factoid questions.)

2.2 List questions

A list question can be thought of as a shorthand for asking the same factoid question multiple times. The set of all correct, distinct answers in the document collection that satisfy the factoid question is the correct answer for the list question.

A system's response for a list question was an unordered set of [*doc-id*, *answer-string*] pairs such that each *answer-string* was considered an instance of the requested type. Judgments of incorrect, unsupported, not exact, and correct were made for individual response pairs as in the factoid judging. The assessor was given one run's entire list at a time, and while judging for correctness also marked a set of responses as distinct. The assessor arbitrarily chose any one of equivalent responses to be distinct, and the remainder were not distinct. Only correct responses could be marked as distinct.

The final set of correct answers for a list question was compiled from the union of the correct responses across all runs plus the instances the assessor found during question development. For the 55 list questions used in the evaluation (one list question was dropped because the assessor decided there were no correct answers during judging), the average

Table 2: Evaluation scores for runs with the best factoid component.

		Accuracy				
Run Tag	Submitter	All	Initial	Non-Initial	NIL Prec	NIL Recall
lcc1	Language Computer Corp.	0.770	0.839	0.744	0.857	0.545
uwbqitekat04	Univ. of Wales, Bangor	0.643	0.694	0.625	0.247	0.864
NUSCHUA1	National Univ. of Singapore	0.626	0.710	0.595	0.333	0.273
mk2004qar1	Saarland University	0.343	0.419	0.315	0.177	0.500
IBM1	IBM Research	0.313	0.435	0.268	—	0.000
mit1	MIT	0.313	0.468	0.256	0.083	0.045
irst04higher	ITC-irst	0.291	0.355	0.268	0.167	0.091
FDUQA13a	Fudan University (Wu)	0.257	0.355	0.220	0.167	0.091
KUQA1	Korea University	0.222	0.226	0.220	0.042	0.045
shef04afv	University of Sheffield	0.213	0.177	0.226	0.071	0.136

Table 3: Average F scores for the list question component. Scores are given for the best run from the top 10 groups.

Run Tag	Submitter	F
lcc1	Language Computer Corp.	0.622
NUSCHUA2	National Univ. of Singapore	0.486
uwbqitekat04	Univ. of Wales, Bangor	0.258
IBM1	IBM Research	0.200
KUQA1	Korea University	0.159
FDUQA13a	Fudan University (Wu)	0.143
MITRE2004B	Mitre Corp.	0.143
UNTQA04M1	University of North Texas	0.128
mk2004qar3	Saarland University	0.125
shef04afv	University of Sheffield	0.125

number of answers per question is 8.8, with 2 as the smallest number of answers, and 41 as the maximum number of answers. A system's response to a list question was scored using instance precision (IP) and instance recall (IR) based on the list of known instances. Let S be the the number of known instances, D be the number of correct, distinct responses returned by the system, and N be the total number of responses returned by the system. Then $IP = D/N$ and $IR = D/S$. Precision and recall were then combined using the F measure with equal weight given to recall and precision:

$$F = \frac{2 \times IP \times IR}{IP + IR}$$

The score for the list component of a run was the average F score over the 55 questions. Table 3 gives the average F scores for the run with the best list component score for each of the top 10 groups.

As happened last year, some submitted runs contained identical list question components as another run submitted by the same group. Since assessors see the lists for each run separately, it can happen that identical components receive different scores. NIST tries to minimize judging differences by making sure the same assessor judges all runs and completes judging one question before moving on to another, but differences remain. These differences are one measure of the error inherent in the evaluation. NIST does not adjust the judgments to make identical runs match because then we wouldn't know what the naturally occurring error rate was, and doing so would bias the scores of systems that submitted identical component runs.

There were 15 pairs of runs with identical list components. Seven pairs had identical average F scores, though some of those seven did have individual questions judged differently. The largest difference in average F scores for identical list components was 0.006, and the largest number of individual questions judged differently for a single run pair was 7.

2.3 Other questions

The Other questions were evaluated using the same methodology as the TREC 2003 definition questions. A system's response for an Other question was an unordered set of [*doc-id*, *answer-string*] pairs as in the list component. Each string was presumed to be a facet in the definition of the series' target that had not yet been covered by earlier questions in the series. The requirement to not repeat information already covered by earlier questions in the series made answering Other questions somewhat more difficult than answering TREC 2003 definition questions.

Judging the quality of the systems' responses was done in two steps. In the first step, all of the answer strings from all of the systems' responses were presented to the assessor in a single list. Using these responses and the searches done during question development, the assessor created a list of information nuggets about the target. An information nugget is an atomic piece of information about the target that is interesting (in the assessor's opinion) and was not part of an earlier question in the series or an answer to an earlier question in the series. An information nugget is atomic if the assessor can make a binary decision as to whether the nugget appears in a response. Once the nugget list was created for a target, the assessor marked some nuggets as vital, meaning that this information must be returned for a response to be good. Non-vital nuggets act as don't care conditions in that the assessor believes the information in the nugget to be interesting enough that returning the information is acceptable in, but not necessary for, a good response.

In the second step of judging the responses, the assessor went through each system's response in turn and marked which nuggets appeared in the response. A response contained a nugget if there was a *conceptual* match between the response and the nugget; that is, the match was independent of the particular wording used in either the nugget or the response. A nugget match was marked at most once per response—if the response contained more than one match for a nugget, an arbitrary match was marked and the remainder were left unmarked. A single [*doc-id*, *answer-string*] pair in a system response could match 0, 1, or multiple nuggets.

Given the nugget list and the set of nuggets matched in a system's response, the nugget recall of the response is the ratio of the number of matched nuggets to the total number of vital nuggets in the list. Nugget precision is much more difficult to compute since there is no effective way of enumerating all the concepts in a response. Instead, a measure based on length (in non-white space characters) is used as an approximation to nugget precision. The length-based measure starts with an initial allowance of 100 characters for each (vital or non-vital) nugget matched. If the total system response is less than this number of characters, the value of the measure is 1.0. Otherwise, the measure's value decreases as the length increases using the function $1 - \frac{length - allowance}{length}$. The final score for an Other question was computed as the F measure with nugget recall three times as important as nugget precision:

$$F(\beta = 3) = \frac{10 \times \text{precision} \times \text{recall}}{9 \times \text{precision} + \text{recall}}.$$

The score for the Other question component was the average F($\beta = 3$) score over 64 Other questions. The Other question for series S7 was mistakenly left unjudged, so it was removed from the evaluation. Table 4 gives the average F($\beta = 3$) score for the best scoring Other question component for each of the top 10 groups.

As with list questions, a system's response for an Other question must be judged as a unit, so identical responses may receive different scores. There were 13 pairs of runs with identical Other question components. The differences between the run pairs' average F($\beta = 3$) scores were {0.012, 0.0, 0.0, 0.0, 0.0, 0.007, 0.0, 0.007, .003, 0.007, 0.0, 0.012, 0.003}, and the number of Other questions that received a different score between the run pairs was {12, 0, 0, 0, 0, 5, 5, 4, 3, 3, 10, 4, 1} respectively.

2.4 Combined weighted average

The final score for a QA run was computed as a weighted average of the three component scores:

$$\text{FinalScore} = .5 \times \text{FactoidAccuracy} + .25 \times \text{ListAveF} + .25 \times \text{OtherAveF}.$$

Since each of the component scores ranges between 0 and 1, the final score is also in that range. Table 5 shows the combined scores for the best run for each of the top 10 groups. Also given in the table are the weighted component scores that make up the final sum.

Table 4: Average F($\beta = 3$) scores for the Other questions component. Scores are given for the best run from the top 10 groups.

Run Tag	Submitter	F($\beta = 3$)
NUSCHUA2	National Univ. of Singapore	0.460
FDUQA13a	Fudan University (Wu)	0.404
NSAQACTIS1	National Security Agency	0.376
ShefMadCow20	University of Sheffield	0.321
UNTQA04M3	University of North Texas	0.307
IBM1	IBM Research	0.285
KUQA3	Korea University	0.247
lcc1	Language Computer Corp.	0.240
clr04r1	CL Research	0.239
mk2004qar3	Saarland University	0.211

Table 5: Weighted component scores and final combined scores for QA task runs. Scores are given for the best run from the top 10 groups.

		Weighted Component Score			Final
Run Tag	Submitter	Factoid	List	Other	Score
lcc1	Language Computer Corp.	0.385	0.155	0.060	0.601
NUSCHUA1	National Univ. of Singapore	0.313	0.120	0.112	0.545
uwbqitekat04	Univ. of Wales, Bangor	0.322	0.065	0.000	0.386
IBM1	IBM Research	0.157	0.050	0.071	0.278
FDUQA13a	Fudan University (Wu)	0.129	0.036	0.101	0.265
mk2004qar3	Saarland University	0.172	0.031	0.053	0.256
mit1	MIT	0.157	0.030	0.046	0.232
irst04higher	ITC-irst	0.145	0.026	0.052	0.223
shef04afv	University of Sheffield	0.106	0.031	0.078	0.216
KUQA1	Korea University	0.111	0.040	0.061	0.212

3 System Approaches

The overall approach taken for answering factoid questions has remained unchanged for the past several years. Systems generally determine the expected answer type of the question, retrieve documents or passages likely to contain answers to the question using important question words and related terms as the query, and then perform a match between the question words and retrieved passages to extract the answer. While the overall approach has remained the same, individual groups continue to refine their techniques for these three steps, increasing the coverage and accuracy of their systems.

Most groups use their factoid-answering system for list questions, changing only the number of responses returned as the answer. The main issue is determining the number of responses to return. Systems whose matching phase creates a question-independent score for each passage return all answers whose score is above an empirically determined threshold. Other systems return all answers whose scores were within an empirically determined fraction of the top result's score.

The fact that target and list questions did not necessarily explicitly include the target of the question was a new difficulty in this year's track. For the document/passage retrieval phase, most systems simply appended the target to the query. This was an effective strategy since in all cases the target was the correct domain for the question, and most of the retrieval methods used treat the query as a simple set of keywords. There were a variety of approaches taken by different systems to address this difficulty in phases that require more detailed processing of the question. While a few systems made no attempt to include the target in the question, a much more common approach was to append the target to the question. Another common approach was to replace all pronouns in the questions with the target. While many (but not all) pronouns in the questions did in fact refer to the target, this approach suffered when the question used a definite noun phrase rather than a pronoun to refer to the target (e.g., using "the band" when the target was Nirvana). Finally, other systems tried varying degrees of true anaphora resolution to appropriately resolve references in the questions. It is difficult to judge how much benefit these systems received from this more extensive processing since the majority of pronoun references were to the target.

Systems generally used the same techniques as were used for TREC 2003's definition questions to answer the Other questions. Most systems first retrieve passages about the target using a recall-oriented retrieval search. Subsequent processing reduces the amount of material returned. Some systems used pattern-matching to locate definition-content in text. These patterns, such as looking for copular constructions and appositives, were either hand-constructed or learned from a training corpus. Systems also looked to eliminate redundant information, using either word overlap measures or document summarization techniques. Unlike last year, answers to previous questions in the series had to be incorporated as part of the redundant information for this year's task. The output from the redundancy-reducing step was then returned as the answer for the Other question.

4 Per-series Combined Weighted Scores

The series play no role in computing the combined average score as above. That is, questions are treated independently without regard to the series they appear in for scoring purposes. This is unfortunate since each individual series is an abstraction of a single user's interaction with the system. Evaluating over the individual series should provide a more accurate representation of the effectiveness of the system from an individual user's perspective. This section examines the effectiveness of a per-series evaluation.

Since each series is a mixture of different question types, we can compute the weighted average score on a per-series basis, and take the average of the per-series scores as the final score for the run. Note that the average per-series weighted score (call this the per-series score) will not in general be equal to the final score computed as the weighted average of the three component scores (the global score) since the two averages emphasize different things. The global score gives equal weight to individual questions within a component. The per-series score gives equal weight to each series. (This is the same difference between micro- and macro-averaging of document retrieval scores.) To compute the combined score for an individual series that contained all three question types, the same weighted average of the different question types was used, but only the scores for questions belonging to the series were part of the computation. For those series that did not contain any list questions, the weighted score was computed as $.67 \times \text{FactoidAccuracy} + .33 \times \text{OtherF}$. All of series S7 was eliminated from the evaluation since that was the series

Table 6: Per-series scores for QA task runs. Scores are given for the best run from the top 10 groups. Also given is the global score (as given in Table 5) for comparison.

Run Tag	Submitter	Per-series	Global
lcc1	Language Computer Corp.	0.609	0.601
NUSCHUA1	National Univ. of Singapore	0.557	0.545
uwbqitekat04	Univ. of Wales, Bangor	0.401	0.386
IBM1	IBM Research	0.289	0.278
FDUQA13a	Fudan University (Wu)	0.289	0.265
mk2004qar3	Saarland University	0.271	0.256
mit1	MIT	0.253	0.232
irst04higher	ITC-irst	0.239	0.223
shef04afv	University of Sheffield	0.230	0.216
NSAQACTIS1	National Security Agency	0.226	0.211

whose Other question was not evaluated.

Table 6 shows the per-series score for the best run for each of the top 10 groups. The global score is repeated in the table for comparison. For the particular set of runs shown in the table, all of the runs rank in the same order by the two scoring methods, except that the tenth run is different for the two schemes (the NSAQACTIS1 run edges out the KUQA1 run when using the per-series score). The absolute value of the per-series score is somewhat greater than the global score for these runs, though it is possible for the global score to be the greater of the two.

Each individual series has only a few questions, so the combined weighted score for an individual series will be much less stable than the global score. But the average of 64 per-series scores should be at least as stable as the overall combined weighted average and has some additional advantages. The per-series score is computed at a small enough granularity to be meaningful at the task-level (i.e., each series representing a single user interaction), and at a large enough granularity for individual scores to be meaningful. Figure 2 shows a box-and-whiskers plot of the per-series scores across all runs for each series. A box in the plot shows the extent of the middle half of the scores for that series, with the median score indicated by the line through the box. The dotted lines (the "whiskers") extend to a point that is 1.5 times the interquartile distance, or the most extreme score, whichever is less. Extreme scores that are greater than the 1.5 times the interquartile distance are plotted as circles. The plot shows that only a few series (S21, S25, S37, S39) have median scores of 0.0. This is in sharp contrast to the median scores of individual questions. For factoid questions, 212 of the 230 questions (92.2%) have a zero median; for list questions 39 of 55 questions (70.9%) have a zero median; and for Other questions 41 of 64 questions (64.1%) have a zero median.

One of the hypotheses during question development was that system effectiveness would depend on the type of target. For example, PERSON targets may be easier for systems to define since the set of information desired for a person may be more standard then the set of information desired for a THING. This hypothesis has little support in the overall results of the track (there may be individual systems that show stronger dependencies). The average of the average per-series score across all runs and all series is 0.187. The averages for series restricted to particular target types are 0.184 for PERSON targets, 0.179 for ORGANIZATION targets, and 0.206 for THING targets.

5 Future of the QA Track

Several concerns regarding the TREC 2005 QA track were raised during the TREC 2004 QA breakout session. Since the TREC 2004 task was rather different from previous years' tasks, there was the desire to repeat the task largely unchanged. There was also the desire to build infrastructure that would allow a closer examination of the role document retrieval techniques play in supporting QA technology. As a result of this discussion, the main task for the 2005 QA track was decided to be essentially the same as the 2004 task in that the test set will consist of a set of question series where each series asks for information regarding a particular target. As in TREC 2004, the targets will include people, organizations, and other entities; unlike TREC 2004 the target can also be an event. Events were added since the document set from which the answers are to be drawn are newswire articles. Each question series will consist of some

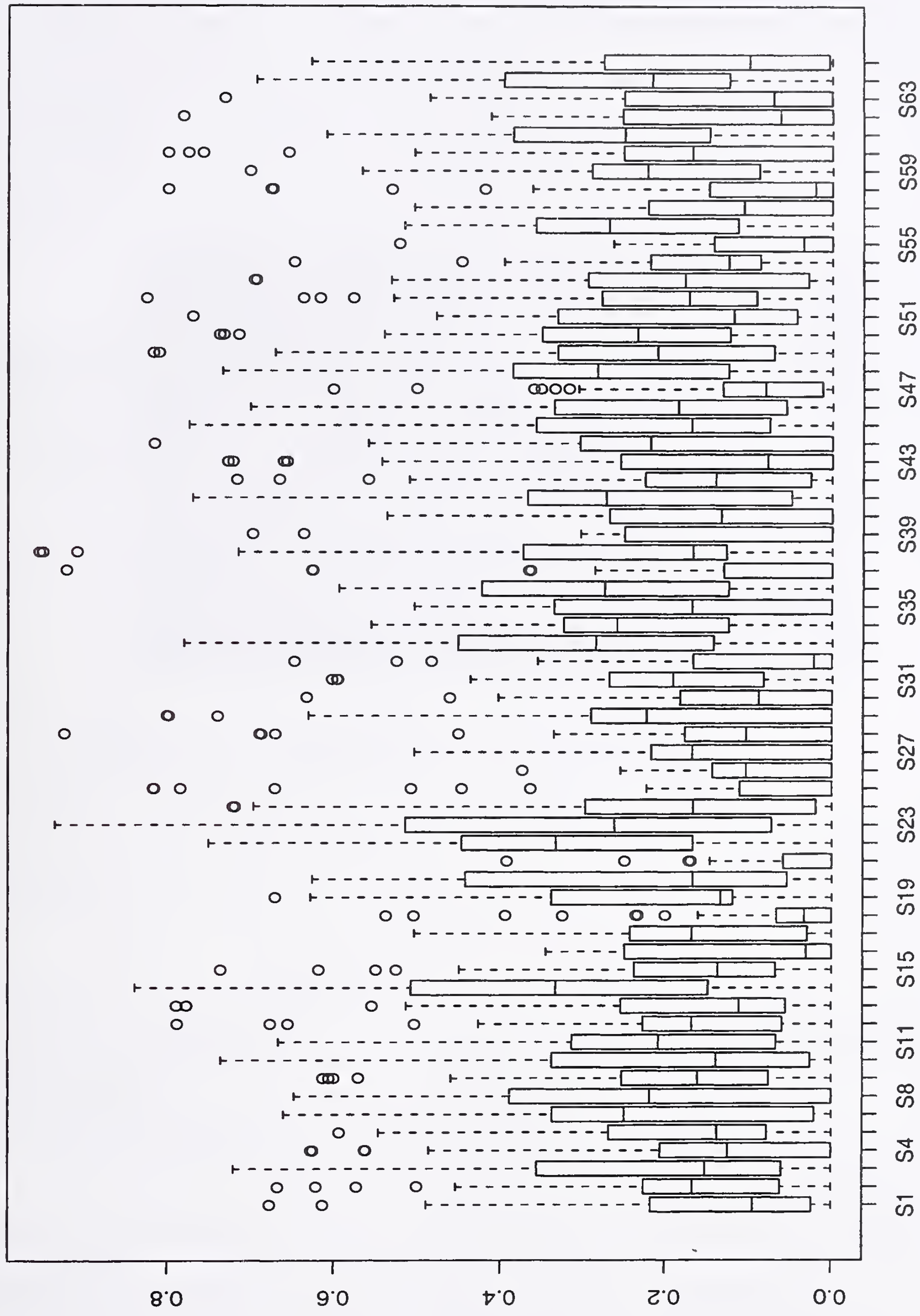


Figure 2: Box and whiskers plot of per-series combined weighted scores across all runs. The x-axis shows the series number (recall that series S7 was omitted), and the y-axis the combined weighted score for that series.

factoid and some list questions and will end with exactly one "Other" question. The answer to the "Other" question is to be interesting information about the target that is not covered by the preceding questions in the series. The runs will be evaluated using the same methodology as in TREC 2004, though the primary measure will be the per-series score.

To address the concern regarding document retrieval and QA, TREC 2005 submissions will be required to include an ordered list of documents for each question. This list will represent the the set of documents used by the system to create its answer, where the order of the documents in the list is the order in which the system considered the document. The purpose of the lists is to create document pools both to get a better understanding of the number of instances of correct answers in the collection and to support research on whether some document retrieval techniques are better than others in support of QA. For some subset of approximately 50 questions, NIST will pool the document lists, and assessors will judge each document in the pool as relevant ("contains an answer") or not relevant ("does not contain an answer"). Document lists will then be evaluated using using trec_eval measures.

References

[1] Tsuneaki Kato, Jun'ichi Fukumoto, Fumito Masui, and Noriko Kando. Handling information access dialogue through QA technologies—A novel challenge for open-domain question answering. In *Proceedings of the HLT-NAACL 2004 Workshop on Pragmatics of Question Answering*, pages 70–77, May 2004.

[2] Ellen M. Voorhees. Overview of the TREC 2001 question answering track. In *Proceedings of the Tenth Text REtreival Conference (TREC 2001)*, pages 42–51, 2002.

[3] Ellen M. Voorhees. Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 54–68, 2004.

Table 5 - Boundary cases for utility measure of triage task for training and test data.

Situation	**U_{norm} - Training**	**U_{norm} - Test**
Completely perfect prediction	1.0	1.0
Triage everything	0.27	0.33
Triage nothing	0	0
Completely imperfect prediction	-0.73	-0.67

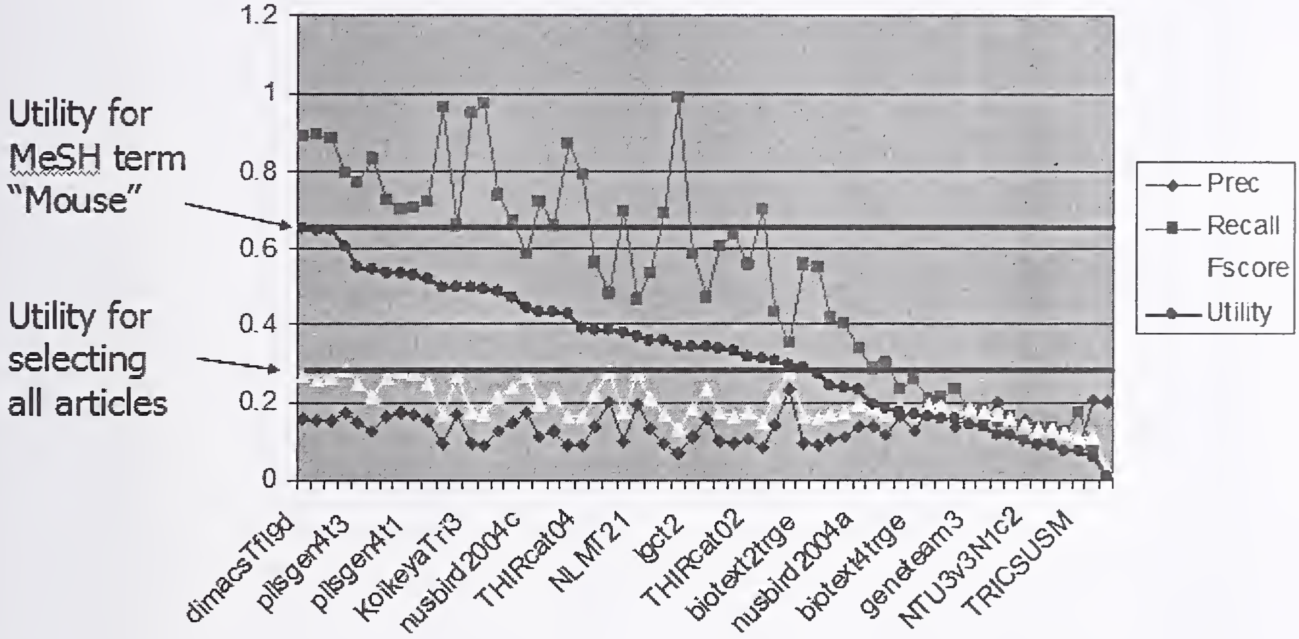


Figure 9 - Triage subtask runs sorted by U_{norm} score. The U_{norm} for the MeSH term *Mice* as well as for selecting all articles as positive is shown.

The evaluation measures for the annotation subtasks were based on the notion of identifying tuples of data. Given the article and gene, systems designated one or both of the following tuples:

- <article, gene, GO hierarchy code>
- <article, gene, GO hierarchy code, evidence code>

We employed a global recall, precision, and F measure evaluation measure for each subtask:

Recall = number of tuples correctly identified / number of correct tuples

Precision = number of tuples correctly identified / number of tuples identified

F = (2 * recall * precision) / (recall + precision)

For the training data, the number of correct <article, gene, GO hierarchy code> tuples was 589, while the number of correct <article, gene, GO hierarchy code, evidence code> tuples was 640.

The annotation hierarchy subtask results are shown in Table 8, while the annotation hierarchy subtask plus evidence code results are shown in Table 9. As noted above, the primary evaluation measure for this task was the F-score. Due to their only being a single measure per run, we were unable to perform comparative statistics. Figure 10 shows the annotation hierarchy subtask results graphically.

Table 6 - Triage subtask runs, sorted by utility.

Run	Group (reference)	Precision	Recall	F-score	Utility
dimacsTfl9d	rutgers.dayanik [16]	0.1579	0.8881	0.2681	0.6512
dimacsTl9mhg	rutgers.dayanik [16]	0.1514	0.8952	0.259	0.6443
dimacsTfl9w	rutgers.dayanik [16]	0.1553	0.8833	0.2642	0.6431
dimacsTl9md	rutgers.dayanik [16]	0.173	0.7952	0.2841	0.6051
pllsgen4t3	patolis.fujita [7]	0.149	0.769	0.2496	0.5494
pllsgen4t4	patolis.fujita [7]	0.1259	0.831	0.2186	0.5424
pllsgen4t2	patolis.fujita [7]	0.1618	0.7238	0.2645	0.5363
pllsgen4t5	patolis.fujita [7]	0.174	0.6976	0.2785	0.532
pllsgen4t1	patolis.fujita [7]	0.1694	0.7024	0.273	0.5302
GUCwdply2000	german.u.cairo [18]	0.151	0.719	0.2496	0.5169
KoikeyaTri1	u.tokyo (none)	0.0938	0.9643	0.171	0.4986
OHSUVP	ohsu.hersh [12]	0.1714	0.6571	0.2719	0.4983
KoikeyaTri3	u.tokyo (none)	0.0955	0.9452	0.1734	0.4974
KoikeyaTri2	u.tokyo (none)	0.0913	0.9738	0.167	0.4893
NLMT2SVM	nlm.umd.ul [15]	0.1286	0.7333	0.2188	0.4849
dimacsTl9w	rutgers.dayanik [16]	0.1456	0.6643	0.2389	0.4694
nusbird2004c	mlg.nus [33]	0.1731	0.5833	0.267	0.444
lgct1	indiana.u.seki [13]	0.1118	0.7214	0.1935	0.4348
OHSUNBAYES	ohsu.hersh [12]	0.129	0.6548	0.2155	0.4337
NLMT2BAYES	nlm.umd.ul [15]	0.0902	0.869	0.1635	0.4308
THIRcat04	tsinghua.ma [9]	0.0908	0.7881	0.1628	0.3935
GUClin1700	german.u.cairo [18]	0.1382	0.5595	0.2217	0.3851
NLMT22	nlm.umd.ul [15]	0.1986	0.481	0.2811	0.3839
NTU2v3N1	ntu.chen [34]	0.1003	0.6905	0.1752	0.381
NLMT21	nlm.umd.ul [15]	0.195	0.4643	0.2746	0.3685
GUCply1700	german.u.cairo [18]	0.1324	0.5357	0.2123	0.3601
NTU3v3N1	ntu.chen [34]	0.0953	0.6857	0.1673	0.3601
NLMT2ADA	nlm.umd.ul [15]	0.0713	0.9881	0.133	0.3448
lgct2	indiana.u.seki [13]	0.1086	0.581	0.183	0.3426
GUClin1260	german.u.cairo [18]	0.1563	0.469	0.2345	0.3425
THIRcat01	tsinghua.ma [9]	0.1021	0.6024	0.1746	0.3375
NTU4v3N1416	ntu.chen [34]	0.0948	0.6357	0.165	0.3323
THIRcat02	tsinghua.ma [9]	0.1033	0.5571	0.1743	0.3154
biotext1trge	u.cberkeley.hearst [14]	0.0831	0.7	0.1486	0.3139
GUCply1260	german.u.cairo [18]	0.1444	0.4333	0.2167	0.305
OHSUSVMJ20	ohsu.hersh [12]	0.2309	0.3524	0.279	0.2937
biotext2trge	u.cberkeley.hearst [14]	0.095	0.5548	0.1622	0.2905
THIRcat03	tsinghua.ma [9]	0.0914	0.55	0.1567	0.2765
THIRcat05	tsinghua.ma [9]	0.1082	0.4167	0.1718	0.245
biotext3trge	u.cberkeley.hearst [14]	0.1096	0.4024	0.1723	0.2389
nusbird2004a	mlg.nus [33]	0.1373	0.3357	0.1949	0.2302
nusbird2004d	mlg.nus [33]	0.1349	0.2881	0.1838	0.1957
nusbird2004b	mlg.nus [33]	0.1163	0.3	0.1677	0.1861
eres2	u.edinburgh.sinclair [32]	0.1647	0.231	0.1923	0.1724
biotext4trge	u.cberkeley.hearst [14]	0.1271	0.2571	0.1701	0.1688
emet2	u.edinburgh.sinclair [32]	0.1847	0.2071	0.1953	0.1614
epub2	u.edinburgh.sinclair [32]	0.1729	0.2095	0.1895	0.1594
nusbird2004e	mlg.nus [33]	0.136	0.231	0.1712	0.1576
geneteam3	u.hospital.geneva [35]	0.1829	0.1833	0.1831	0.1424
edis2	u.edinburgh.sinclair [32]	0.1602	0.1857	0.172	0.137
wdtriage1	indiana.u.yang [27]	0.202	0.1476	0.1706	0.1185
eint2	u.edinburgh.sinclair [32]	0.1538	0.1619	0.1578	0.1174
NTU3v3N1c2	ntu.chen [34]	0.1553	0.1357	0.1449	0.0988
geneteam1	u.hospital.geneva [35]	0.1333	0.1333	0.1333	0.09
geneteam2	u.hospital.geneva [35]	0.1333	0.1333	0.1333	0.09
biotext5trge	u.cberkeley.hearst [14]	0.1192	0.1214	0.1203	0.0765
TRICSUSM	u.sanmarcos [31]	0.0792	0.1762	0.1093	0.0738
IBMIRLver1	ibm.india (none)	0.2053	0.0738	0.1086	0.0595
EMCTNOT1	tno.kraaij [19]	0.2	0.0143	0.0267	0.0114
Mean		0.1381	0.5194	0.1946	0.3303
MeSH *Mice*	rutgers.dayanik [16]	0.1502	0.8929	0.2572	0.6404

Table 7 - Data file contents and counts for annotation hierarchy subtasks.

File contents	Training data count	Test data count
Documents - PMIDs	504	378
Genes - Gene symbol, MGI identifier, and gene name for all used	1294	777
Document gene pairs - PMID-gene pairs	1418	877
Positive examples - PMIDs	178	149
Positive examples - PMID-gene pairs	346	295
Positive examples - PMID-gene-domain tuples	589	495
Positive examples - PMID-gene-domain-evidence tuples	640	522
Positive examples - all PMID-gene-GO-evidence tuples	872	693
Negative examples - PMIDs	326	229
Negative examples - PMID-gene pairs	1072	582

Table 8 - Annotation hierarchy subtask, sorted by F-score.

Run	Group (reference)	Precision	Recall	F-score
lgcad1	indiana.u.seki [13]	0.4415	0.7697	0.5611
lgcad2	indiana.u.seki [13]	0.4275	0.7859	0.5537
wiscWRT	u.wisconsin [17]	0.4386	0.6202	0.5138
wiscWT	u.wisconsin [17]	0.4218	0.6263	0.5041
dimacsAg3mh	rutgers.dayanik [16]	0.5344	0.4545	0.4913
NLMA1	nlm.umd.ul [15]	0.4306	0.5515	0.4836
wiscWR	u.wisconsin [17]	0.4255	0.5596	0.4834
NLMA2	nlm.umd.ul [15]	0.427	0.5374	0.4758
wiscW	u.wisconsin [17]	0.3935	0.5596	0.4621
KoikeyaHi1	u.tokyo (none)	0.3178	0.7293	0.4427
iowarun3	u.iowa [23]	0.3207	0.6	0.418
iowarun1	u.iowa [23]	0.3371	0.5434	0.4161
iowarun2	u.iowa [23]	0.3812	0.4505	0.413
BIOTEXT22	u.cberkeley.hearst [14]	0.2708	0.796	0.4041
BIOTEXT21	u.cberkeley.hearst [14]	0.2658	0.8141	0.4008
dimacsAl3w	rutgers.dayanik [16]	0.5015	0.3273	0.3961
GUCsvm0	german.u.cairo [18]	0.2372	0.7414	0.3595
GUCir50	german.u.cairo [18]	0.2303	0.8081	0.3584
geneteamA5	u.hospital.geneva [35]	0.2274	0.7859	0.3527
GUCir30	german.u.cairo [18]	0.2212	0.8404	0.3502
geneteamA4	u.hospital.geneva [35]	0.209	0.9354	0.3417
BIOTEXT24	u.cberkeley.hearst [14]	0.4452	0.2707	0.3367
GUCsvm5	german.u.cairo [18]	0.2052	0.9354	0.3366
cuhkrun3	chinese.u.hongkong (none)	0.4174	0.2808	0.3357
geneteamA2	u.hospital.geneva [35]	0.2025	0.9535	0.334
dimacsAabsw1	rutgers.dayanik [16]	0.5979	0.2283	0.3304
BIOTEXT23	u.cberkeley.hearst [14]	0.4437	0.2626	0.3299
geneteamA1	u.hospital.geneva [35]	0.1948	0.9778	0.3248
geneteamA3	u.hospital.geneva [35]	0.1938	0.9798	0.3235
GUCbase	german.u.cairo [18]	0.1881	1	0.3167
BIOTEXT25	u.cberkeley.hearst [14]	0.4181	0.2525	0.3149
cuhkrun2	chinese.u.hongkong (none)	0.4385	0.2303	0.302
cuhkrun1	chinese.u.hongkong (none)	0.4431	0.2283	0.3013
dimacsAp5w5	rutgers.dayanik [16]	0.5424	0.1939	0.2857
dimacsAw20w5	rutgers.dayanik [16]	0.6014	0.1677	0.2622
iowarun4	u.iowa [23]	0.1692	0.1333	0.1492
Mean		0.3600	0.5814	0.3824

Table 9 - Annotation hierarchy plus evidence code subtask, sorted by F-score.

Tag	Group (reference)	Precision	Recall	F-score
lgcab2	indiana.u.seki [13]	0.3238	0.6073	0.4224
lgcab1	indiana.u.seki [13]	0.3413	0.4923	0.4031
KoikeyaHiev1	u.tokyo (none)	0.2025	0.4406	0.2774
Mean		0.2892	0.5134	0.3676

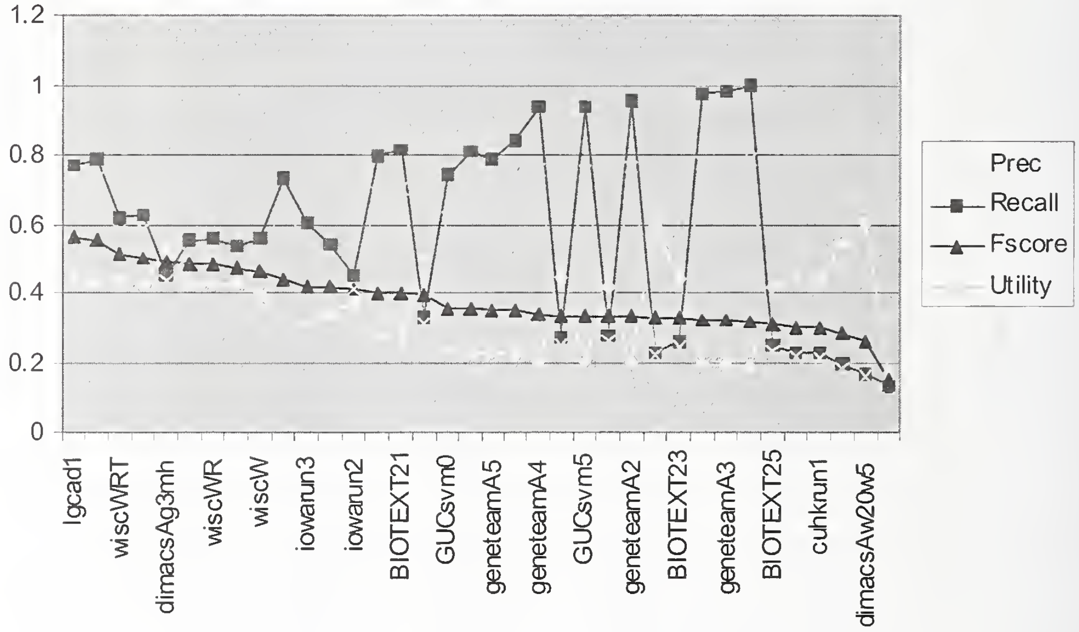


Figure 10 - Annotation hierarchy subtask results sorted by F-score.

In the annotation hierarchy subtask, the runs varied widely in recall and precision. The best runs, i.e., those with the highest F-scores, had medium levels of recall and precision. The top run came from Indiana University and used a variety of approaches, including a k-nearest neighbor model, mapping terms to MeSH, using keyword and glossary fields of documents, and recognizing gene names [13]. Further post-submission runs raised their F-score to 0.639. Across a number of groups, benefit was found from matching gene names appropriately. University of Wisconsin also found identifying gene names in sentences and modeling features in those sentences provided value [17].

5. Discussion

The TREC 2004 Genomics Track was very successful, with a great deal of enthusiastic participation. In all of the tasks, a diversity of approaches were used, resulting in wide variation across the results. Trying to discern the relative value of them is challenging, since few groups performed parameterized experiments or used common baselines.

In the ad hoc retrieval task, the best approaches employed techniques known to be effective in non-biomedical TREC tasks. These included Okapi weighting, blind relevance feedback, and language modeling. However, some domain-specific approaches appeared to be beneficial, such as expanding queries with synonyms from controlled vocabularies that are widely available. There also appeared to be some benefit for boosting parts of the queries. However, it was also easy for many groups to do detrimental things, as evidenced by the OHSU run of a TF*IDF system "out of the box" that scored well above the median.

The triage subtask was limited by the fact that using the MeSH term *Mice* assigned by the MEDLINE indexers was a better predictor of the MGI triage decision than anything else, including the complex feature extraction and machine learning algorithms of many participating groups. Some expressed concern that MGI might give preference to basing annotation decisions on maximizing coverage of genes instead of exhaustively cataloging the literature, something that would be useful for users of its system but compromise the value of its data in tasks like automated article triage. We were assured by the MGI director (J. Blake, personal communication) that the initial triage decision for an article was made independent of the prior coverage of gene, even though priority decisions made later in the pipeline did take coverage into account. As such, the triage decision upon which our data were based was sound from the standpoint of document classification. The annotation decision was also not effected by this since the positive and negative are not exhaustive (and do not need to be) for this subtask.

Another concern about the MGI data was whether the snapshot obtained in mid-2004 was significantly updated by the time the track was completed. This was analyzed in early 2005, and it was indeed found that the number of PMIDs in the triage subtask had increased in size by about 10%, with a very small number now negatively triaged. While this change is unlikely to have major impact on results, an updated data set will be released in early 2005.

But the remaining question for the triage subtask is why systems were unable to outperform the MeSH term *Mice*. It should be noted that this term was far from perfect, achieving a recall of 89% but a precision of only 15%. So why cannot more elaborate systems outperform this? There are a variety of explanations:

- MGI data is problematic - while MGI does some internal quality checking, they do not carry it out at the level that research groups would, e.g., with kappa scores
- Our algorithms and systems are imperfect - we do not know or there do not exist better predictive features
- Our metrics may be problematic - is the factor = 20 in the utility formula appropriate?

We believe that the triage subtask data represents an important task (i.e., document triage is valuable in a variety of biomedical settings, such as discerning the best evidence in clinical studies) and that these data provide the substrate for work to continue in this area.

The annotation hierarchy task had lower participation, and the value of picking the correct hierarchy is unclear. However, there would be great value to systems that could perform automated GO annotation, even though the task is very challenging [2]. These results demonstrated a value identifying gene names and other controlled vocabulary terms in documents for this task.

The TREC Genomics Track will be continuing in 2005. In addition, the data for the 2004 track will be released to the general community for continued experimentation. The categorization task data will be updated before its release, and both the old and new data will be made available. We hope that all of this will continue to facilitate in IR in the genomics domain.

Acknowledgements

The TREC 2004 Genomics Track was supported by NSF Grant ITR-0325160. The track also appreciates the help of Ellen Voorhees and NIST.

References

1. Mobasheri A, et al., *Post-genomic applications of tissue microarrays: basic research, prognostic oncology, clinical genomics and drug discovery.* Histology and Histopathology, 2004. 19: 325-335.
2. Hirschman L, et al., *Accomplishments and challenges in literature data mining for biology.* Bioinformatics, 2002. 18: 1553-1561.
3. Anonymous, *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Research, 2004. 32: D258-D261.
4. Hersh WR and Bhupatiraju RT. *TREC genomics track overview. The Twelfth Text Retrieval Conference: TREC 2003.* 2003. Gaithersburg, MD: National Institute of Standards and Technology. 14-23. http://trec.nist.gov/pubs/trec12/papers/GENOMICS.OVERVIEW3.pdf.
5. Kramer MS and Feinstein AR, *Clinical biostatistics: LIV. The biostatistics of concordance.* Clinical Pharmacology and Therapeutics, 1981. 29: 111-123.
6. Hersh WR, et al. *OHSUMED: an interactive retrieval evaluation and new large test collection for research. Proceedings of the 17th Annual International ACM SIGIR Conference on*

Research and Development in Information Retrieval. 1994. Dublin, Ireland: Springer-Verlag. 192-201.

7. Fujita S. *Revisiting again document length hypotheses - TREC 2004 Genomics Track experiments at Patolis. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/patolis.geo.pdf.
8. Buttcher S, Clarke CLA, and Cormack GV. *Domain-specific synonym expansion and validation for biomedical information retrieval (MultiText experiments for TREC 2004). The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/uwaterloo-clarke.geo.pdf.
9. Li J, et al. *THUIR at TREC 2004: Genomics Track. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/tsinghua-ma.geo.pdf.
10. Carpenter B. *Phrasal queries with LingPipe and Lucene: ad hoc genomics text retrieval. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/alias-i.geo.pdf.
11. Pirkola A. *TREC 2004 Genomics Track experiments at UTA: the effects of primary keys, bigram phrases and query expansion on retrieval performance. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/utampere.geo.pdf.
12. Cohen AM, Bhuptiraju RT, and Hersh W. *Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/ohsu-hersh.geo.pdf.
13. Seki K, et al. *TREC 2004 Genomics Track experiments at IUB. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/indianau-seki.geo.pdf.
14. Nakov PI, et al. *BioText team experiments for the TREC 2004 Genomics Track. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/ucal-berkeley.geo.pdf.
15. Aronson AR, et al. *Knowledge-intensive and statistical approaches to the retrieval and annotation of genomics MEDLINE citations. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/nlm-umd-ul.geo.pdf.
16. Dayanik A, et al. *DIMACS at the TREC 2004 Genomics Track. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/rutgers-dayanik.geo.pdf.
17. Settles B and Craven M. *Exploiting zone information, syntactic rules, and informative terms in Gene Ontology annotation of biomedical documents. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/uwisconsin.geo.pdf.
18. Darwish K and Madkour A. *The GUC goes to TREC 2004: using whole or partial documents for retrieval and classification in the Genomics Track. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/german.u.geo.pdf.
19. Kraaij W, et al. *MeSH based feedback, concept recognition and stacked classification for curation tasks. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/tno-emc.geo.pdf.
20. Crangle C, et al. *Concept extraction and synonymy management for biomedical information retrieval. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of

Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/conversp eech.geo.pdf.

21. Billerbeck B, et al. *RMIT University at TREC 2004. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/rmit.tera.g eo.pdf.
22. Tong RM. *Information needs and automatic queries. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/tarragon.t ong.geo.pdf.
23. Eichmann D, et al. *Novelty, question answering and genomics: the University of Iowa response. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/uiowa.nov elty.qa.geo.pdf.
24. Bacchin M and Melucci M. *Expanding queries using stems and symbols. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/upadova.g eo.pdf.
25. Ruiz ME, Srikanth M, and Srihari R. *UB at TREC 13: Genomics Track. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/stateuny-buffalo.geo.pdf.
26. Huang X, et al. *York University at TREC 2004: HARD and Genomics Tracks. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/yorku.har d.geo.pdf.
27. Yang K, et al. *WIDIT in TREC 2004 Genomics, Hard, Robust and Web Tracks. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/indianau.g eo.hard.robust.web.pdf.
28. Blott S, et al. *Experiments in terabyte searching, genomic retrieval and novelty detection for TREC 2004. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/dcu.tera.g eo.novelty.pdf.
29. Guo Y, Harkema H, and Gaizauskas R. *Sheffield University and the TREC 2004 Genomics Track: query expansion using synonymous terms. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/usheffield. geo.pdf.
30. Tomiyama T, et al. *Meiji University Web, Novelty and Genomic Track experiments. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/meijiu.we b.novelty.geo.pdf.
31. Guillen R. *Categorization of genomics text based on decision rules. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/cal-state-sanmarcos.geo.pdf.
32. Sinclair G and Webber B. *TREC Genomics 2004. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/uedinburg h-sinclair.geo.pdf.
33. Zhang D and Lee WS. *Experience of using SVM for the triage task in TREC 2004 Genomics Track. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/natusing.z hang.geo.pdf.
34. Lee C, Hou WJ, and Chen HH. *Identifying relevant full-text articles for GO annotation without MeSH terms. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/ntu.geo.pd f.
35. Ruch P, et al. *Report on the TREC 2004 experiment: Genomics Track. The Thirteenth Text Retrieval Conference: TREC 2004*. 2004. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec13/papers/uhosp-geneva.geo.pdf.

Overview of the TREC 2004 Robust Retrieval Track

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

The robust retrieval track explores methods for improving the consistency of retrieval technology by focusing on poorly performing topics. The retrieval task in the track is a traditional ad hoc retrieval task where the evaluation methodology emphasizes a system's least effective topics. The most promising approach to improving poorly performing topics is exploiting text collections other than the target collection such as the web.

The 2004 edition of the track used 250 topics and required systems to rank the topics by predicted difficulty. The 250 topics within the test set allowed the stability of evaluation measures that emphasize poorly performing topics to be investigated. A new measure, a variant of the traditional MAP measure that uses a geometric mean rather than an arithmetic mean to average individual topic results, shows promise of giving appropriate emphasis to poorly performing topics while being more stable at equal topic set sizes.

The ability to return at least passable results for any topic is an important feature of an operational retrieval system. While system effectiveness is generally reported as average effectiveness, an individual user does not see the average performance of the system, but only the effectiveness of the system on his or her requests. A user whose request retrieves nothing of interest is unlikely to be consoled by the fact that the system responds better to other people's requests.

The TREC robust retrieval track was started in TREC 2003 to investigate methods for improving the consistency of retrieval technology. The first year of the track had two main technical results:

1. The track provided ample evidence that optimizing average effectiveness using the standard Cranfield methodology and standard evaluation measures further improves the effectiveness of the already-effective topics, sometimes at the expense of the poor performers.

2. The track results demonstrated that measuring poor performance is intrinsically difficult because there is so little signal in the sea of noise for a poorly performing topic. Two new measures devised to emphasize poor performers did so, but because there is so little information the measures are unstable. Having confidence in the conclusion that one system is better than another using these measures requires larger differences in scores than are generally observed in practice when using 50 topics.

The retrieval task in the track is a traditional ad hoc task. In addition to calculating scores using `trec_eval`, each run is also evaluated using the two measures introduced in the TREC 2003 track that focus more specifically on the least-well-performing topics. The TREC 2004 track differed from the initial track in two important ways. First, the test set of topics consisted of 249 topics, up from 100 topics. Second, systems were required to rank the *topics* by predicted difficulty, with the goal of eventually being able to use such predictions to do topic-specific processing.

This paper presents an overview of the results of the track. The first section describes the data used in the track, and the following section gives the retrieval results. Section 3 investigates how accurately systems can predict which topics are difficult. Since one of the main results of the TREC 2003 edition of the track was that the poor performance is hard to measure with 50 topics, section 4 examines the stability of the evaluation measures for larger topic set sizes. The final section looks at the future of the track.

1 The Robust Retrieval Task

As mentioned, the task within the robust retrieval track is a traditional ad hoc task. Since the TREC 2003 track had shown that 50 topics was not sufficient for a stable evaluation of poorly performing topics, the TREC 2004 track used

Table 1: Relevant document statistics for topic sets.

Topic Set	Number of topics	Mean Relevant per Topic	Minimum # Relevant	Maximum # Relevant
Old	200	76.8	3	448
New	49	42.1	3	161
Hard	50	88.3	5	361
Combined	249	69.9	3	448

a set of 250 topics (one of which was subsequently dropped due to having no relevant documents). The topic set consisted of 200 topics that had been used in some prior TREC plus 50 topics created for this year's track. The 200 old topics were the combined set of topics used in the ad hoc task in TRECs 6–8 (topics 301–450) plus the topics developed for the TREC 2003 robust track (topics 601–650). The 50 new topics created for this year's track are topics 651–700. The document collection was the set of documents on TREC disks 4 and 5, minus the *Congressional Record*, since that was the document set used with the old topics in the previous TREC tasks. This document set contains approximately 528,000 documents and 1,904 MB of text.

In the TREC 2003 robust track, 50 of the topics from the 301–450 set were distinguished as being particularly difficult for retrieval systems. These topics each had low median average precision scores but at least one high outlier score in the initial TREC in which they were used. Effectiveness scores over this topic set remained low in the 2003 robust track. This topic set is designated as the "hard" set in the discussion below.

While using old topics allows the test set to contain many topics with at least some of the topics known to be difficult, it also means that full relevance data for these topics is available to the participants. Since we could not control how the old topics had been used in the past, the assumption was that the old topics were fully exploited in any way desired in the construction of a participants' retrieval system. In other words, participants were allowed to explicitly train on the old topics if they desired to. The only restriction placed on the use of relevance data for the old topics was that the relevance judgments could not be used during the processing of the submitted runs. This precluded such things as true (rather than pseudo) relevance feedback and computing weights based on the known relevant set.

The existing relevance judgments were used for the old topics; no new judgments of any kind were made for these topics. The new topics were judged by creating pools from three runs per group and using the top 100 documents per run. There was an average of 704 documents judged for each new topic. The assessors made three-way judgments of not relevant, relevant, or highly relevant for the new topics. As noted above, topic 672 had no documents judged relevant for it, so it was dropped from the evaluation. An additional 10 topics had no documents judged highly relevant. All the evaluation results reported for the track consider both relevant and highly relevant documents as the relevant set. Table 1 gives the total number of topics, the average number of relevant documents, and the minimum and maximum number of relevant documents for a topic for the four topic sets used in the track.

While no new judgments were made for the old topics, NIST did form pools for those topics to examine the coverage of the original judgment set. Across the set of 200 old topics, an average of 70.8% (minimum 36.6%, maximum 93.7%) of the documents in the pools created using robust track runs were judged. Across the 110 runs that were submitted to the track, there was an average of 0.3 (min 0.0, max 2.9) unjudged documents in the top 10 documents retrieved, and 11.2 (min 2.9, max 37.5) unjudged documents in the top 100 retrieved. The runs with the largest number of unjudged documents were also the runs that performed the least well. This make sense in that the irrelevant documents retrieved by these runs are unlikely to be in the the original judgment set. While it is possible that the runs were scored as being ineffective *because* they had large numbers of unjudged documents, this is unlikely to be the case since the same runs were ineffective when evaluated over just the new set of topics.

Runs were evaluated using `trec_eval`, with average scores computed over the set of 200 old topics, the set of 49 new topics, the set of 50 hard topics, and the combined set of 249 topics. Two additional measures that were introduced in the TREC 2003 track were computed over the same four topic sets [11]. The *%no* measure is the percentage of topics that retrieved no relevant documents in the top ten retrieved. The *area* measure is the area under the curve produced by plotting MAP(X) vs. X when X ranges over the worst quarter topics. Note that since the area measure is computed over the individual system's worst X topics, different systems' scores are computed over a different set of topics in general.

Table 2: Groups participating in the robust track.

Chinese Academy of Sciences (CAS-NLPR)	Fondazione Ugo Bordoni
Hong Kong Polytechnic University	Hummingbird
IBM Research, Haifa	Indiana University
Johns Hopkins University/APL	Max-Planck Institute for Computer Science
Peking University	Queens College, CUNY
Sabir Research, Inc.	University of Glasgow
University of Illinois at Chicago	Virginia Tech

2 Retrieval Results

The robust track received a total of 110 runs from the 14 groups listed in Table 2. All of the runs submitted to the track were automatic runs, (most likely because there were 250 topics in the test set). Participants were allowed to submit up to 10 runs. To have comparable runs across participating sites, one run was required to use just the description field of the topic statements, one run was required to use just the title field of the topic statements, and the remaining runs could use any combination of fields. There were 31 title-only runs and 32 description-only runs submitted to the track. There was a noticeable difference in effectiveness depending on the portion of the topic statement used: runs using both the title and description fields were better than using either field in isolation.

Table 3 gives the evaluation scores for the best run for the top 10 groups who submitted either a title-only run or a description-only run. The table gives the scores for the four main measures used in the track as computed over the old topics only, the new topics only, the difficult topics, and all 249 topics. The four measures are mean average precision (MAP), the average of precision at 10 documents retrieved (P10), the percentage of topics with no relevant in the top 10 retrieved (%no), and the area underneath the MAP(X) vs. X curve (area). The run shown in the table is the run with the highest MAP score as computed over the combined topic set; the table is sorted by this same value.

2.1 Retrieval methods

All of the top-performing runs used the web to expand queries [5, 6, 1]. In particular, Kwok and his colleagues had the most effective runs in both TREC 2003 and 2004 by treating the web as a large, domain-independent thesaurus and supplementing the topic statement by its terms [5]. When performed carefully, query expansion by terms in a collection other than the target collection can increase the effectiveness of many topics, including poorly performing topics. Expansion based on the target collection does not help the poor performers because pseudo-relevance feedback needs some relevant documents in the top retrieved to be effective, and that is precisely what the poorly performing topics don't have. The web is not a panacea, however, in that some approaches to exploiting the web can be more harmful than helpful [14].

Other approaches to improving the effectiveness of poor performers included selecting a query processing strategy based on a prediction of topic effectiveness[15, 8], and reodering the original ranking in a post-retrieval phase [7, 13]. Weighting functions, topic fields, and query expansion parameters were selected depending upon the prediction of topic difficulty. Documents were reordered based on trying to ensure different aspects of the topic were all represented. While each of these techniques can help some topics, the improvement was not as consistent as expanding by an external corpus.

2.2 Difficult topics

One obvious aspect of the results is that the hard topics remain hard. Evaluation scores when computed over just the hard topics are approximately half as good as they are when computed over all topics for all measures except P(10) which doesn't degrade quite as badly. While the robust track results don't say anything about why these topics are hard, the 2003 NRRC RIA workshop [4] performed failure analysis on 45 topics from the 301–450 topic set. As one of the results of the failure analysis, Buckley assigned each of the 45 topics into 10 failure categories [2]. He ordered the categories by the amount of natural language understanding (NLU) he thought would be required to get good

Table 3: Evaluation results for the best title-only run (a), and best description-only run (b) for the top 10 groups as measured by MAP over the combined topic set. Runs are ordered by MAP over the combined topic set. Values given are the mean average precision (MAP), precision at rank 10 averaged over topics (P10), the percentage of topics with no relevant in the top ten retrieved (%no), and the area underneath the MAP(X) vs. X curve (area) as computed for the set of 200 old topics, the set of 49 new topics, the set of 50 hard topics, and the combined set of 249 topics.

	Old Topic Set				New Topic Set				Hard Topic Set				Combined Topic Set			
Tag	MAP	P10	%no	area	MAP	P10	%no	area	MAP	P10	%no	area	MAP	P10	%no	area
pircRB04t3	0.317	0.505	5	0.033	0.401	0.545	6	0.089	0.183	0.374	12	0.016	0.333	0.513	5	0.038
fub04Tge	0.298	0.484	13	0.019	0.351	0.480	12	0.046	0.145	0.338	22	0.008	0.309	0.483	12	0.021
uic0401	0.305	0.490	5	0.026	0.325	0.441	6	0.047	0.194	0.376	4	0.026	0.309	0.480	5	0.028
uogRobSWR10	0.296	0.461	16	0.010	0.322	0.453	12	0.021	0.136	0.316	26	0.003	0.301	0.459	15	0.011
vtumtitle	0.278	0.440	20	0.007	0.299	0.429	14	0.015	0.136	0.272	36	0.001	0.282	0.437	19	0.008
humR04t5e1	0.272	0.462	13	0.016	0.298	0.457	12	0.029	0.136	0.332	20	0.009	0.277	0.461	13	0.017
JuruTitSwQE	0.255	0.443	10	0.017	0.271	0.412	10	0.019	0.116	0.282	12	0.009	0.258	0.437	10	0.017
SABIR04BT	0.244	0.416	18	0.008	0.290	0.392	20	0.010	0.115	0.238	32	0.002	0.253	0.411	18	0.008
apl04rsTs	0.239	0.408	13	0.013	0.270	0.386	10	0.021	0.113	0.264	14	0.009	0.245	0.404	12	0.014
polyutp3	0.225	0.420	14	0.006	0.255	0.388	10	0.019	0.083	0.244	24	0.002	0.231	0.414	13	0.007

(a) title-only runs

pircRB04d4	0.316	0.507	8	0.023	0.407	0.547	2	0.074	0.162	0.382	12	0.013	0.334	0.515	7	0.028
fub04Dge	0.309	0.508	9	0.025	0.382	0.535	8	0.044	0.147	0.336	18	0.017	0.324	0.513	9	0.027
uogRobDWR10	0.286	0.454	16	0.007	0.374	0.529	12	0.023	0.131	0.296	28	0.002	0.303	0.468	15	0.008
vtumdesc	0.283	0.449	15	0.007	0.340	0.478	12	0.021	0.132	0.304	20	0.005	0.294	0.455	14	0.008
humR04d4e5	0.265	0.436	18	0.008	0.320	0.480	16	0.023	0.140	0.340	20	0.007	0.276	0.445	17	0.009
JuruDesQE	0.266	0.466	11	0.010	0.295	0.398	16	0.022	0.152	0.348	14	0.008	0.272	0.452	12	0.011
SABIR04BD	0.243	0.429	18	0.007	0.342	0.488	10	0.033	0.114	0.276	32	0.003	0.263	0.441	16	0.009
wdoqdn1	0.248	0.461	10	0.016	0.262	0.412	10	0.028	0.126	0.322	18	0.010	0.251	0.451	10	0.017
apl04rsDw	0.192	0.351	15	0.007	0.237	0.363	8	0.022	0.107	0.264	16	0.005	0.201	0.353	13	0.008
polyudp2	0.185	0.364	16	0.003	0.234	0.378	6	0.025	0.083	0.240	24	0.001	0.195	0.367	14	0.004

(b) description-only runs

effectiveness for the topics in that category, and suggested that topics in categories 1–5 should be amenable to today's technology if systems could detect what category the topic was in. More than half of the 45 topics studied during RIA were placed in the first 5 categories.

Twenty-six topics are in the intersection of the robust track's hard set and the RIA failure analysis set. Table 4 shows how the topics in the intersection were categorized by Buckley. Seventeen of the 26 topics in the intersection are in the earlier categories, suggesting that the hard topic set should not be a hopelessly difficult topic set.

3 Predicting difficulty

A necessary first step in determining the problem with a topic is the ability to recognize whether or not it will be effective. Obviously, to be useful the system needs to be able to make this determination at run time and without any explicit relevance information. Cronen-Townsend, Zhou, and Croft suggested the *clarity measure*, the relative entropy between a query language model and the corresponding collection language model, as one way of predicting the effectiveness of a query [3]. The robust track required systems to rank the topics in the test set by predicted difficulty to explore how capable systems are at recognizing difficult topics. A similar investigation in the TREC 2002 question answering track demonstrated that accurately predicting whether a correct answer was retrieved is a challenging problem [10].

In addition to including the retrieval results for each topic, a robust track run ranked the topics in strict order from 1 to 250 such that the topic at rank 1 was the topic the system predicted it had done best on, the topic at rank 2 was the topic the system predicted it had done next best on, etc. This ranking was the *predicted* ranking. Once the evaluation was complete, the topics were ranked from best to worst by average precision score; this ranking was the

Table 4: Failure categories of hard topics.

Category number	Category gloss	Topics
2	general technical failures such as stemming	353, 378
3	systems all emphasize one aspect, miss another required term	322, 419, 445
4	systems all emphasize one aspect, miss another aspect	350, 355, 372, 408, 409, 435, 443
5	some systems emphasize one aspect, some another, need both	307, 310, 330, 363, 436
6	systems all emphasize some irrelevant aspect, missing point of topic	347
7	need outside expansion of "general" term (e.g., expand Europe to individual countries)	401, 443, 448
8	need query analysis to determine relationship between query terms	414
9	systems missed difficult aspect	362, 367, 389, 393, 401, 404

actual ranking.

One measure for how well two rankings agree is Kendall's τ [9]. Kendall's τ measures the similarity between two rankings as a function of the number of pairwise swaps needed to turn one ranking into the other. The τ ranges between -1.0 and 1.0 where the expected correlation between two randomly generated rankings is 0.0, and a τ of 1.0 indicates perfect agreement. The run with the largest τ between the predicted and actual ranking was the `uic0401` run with a τ of 0.623. Fourteen of the 110 runs submitted to the track had a negative correlation between the predicted and actual rankings. (The topic that was dropped from the evaluation was also removed from the rankings before the τ was computed.)

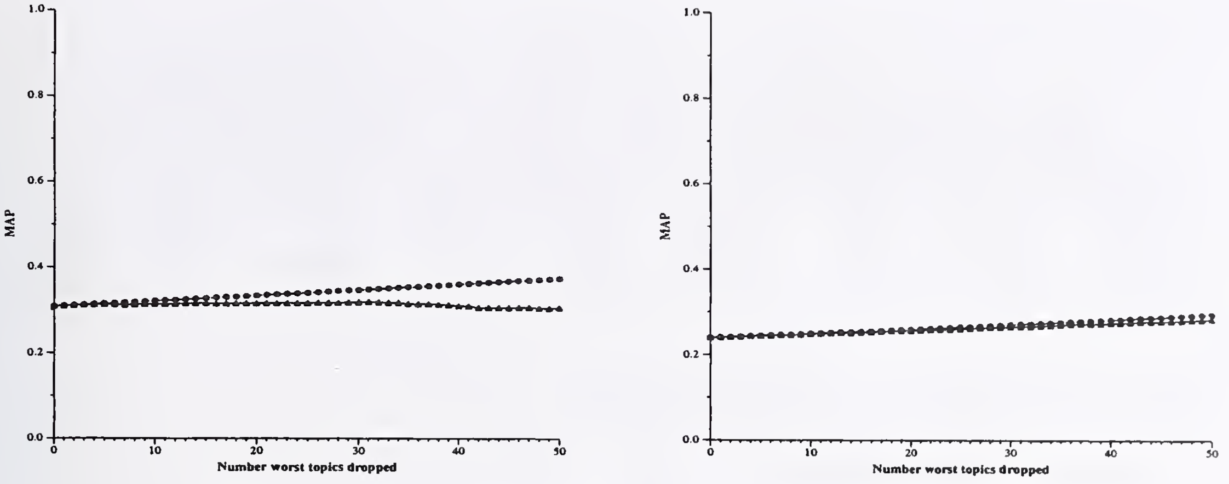
The Kendall's τ score between the predicted and actual ranking for a run is given as part of the run's description in the Appendix of these proceedings. Unfortunately, Kendall's τ between the entire predicted and actual rankings is not a very good measure of whether a system can recognize poorly performing topics. The main problem is that Kendall's τ is sensitive to any difference in the rankings (by design). But for the purposes of predicting when a topic will be a poor performer, small differences in average precision don't matter, nor does the actual ranking of the very effective topics.

A more accurate representation of how well systems predict poorly performing topics is to look at how MAP scores change when successively greater numbers of topics are eliminated from the evaluation. The idea is essentially the inverse of the area measure: instead of computing MAP over the X worst topics, compute it over the best Y topics where $Y = 249 \ldots 199$ and the best topics are defined as the first Y topics in either the predicted or actual ranking. The difference between the two curves produced using the actual ranking on the one hand and the predicted ranking on the other is the measure of how accurate the predictions are. Figure 1 shows these curves plotted for the `uic0401` run, the run with the highest Kendall correlation, on the left and the `humR04d5` run, the run with the (second[1]) smallest difference between curves, on the right. In the figure, the MAP scores computed when eliminating topics from the actual ranking are plotted with circles and scores using the predicted ranking are plotted with triangles.

Figure 2 shows a scatter plot of the area between the MAP curves versus the Kendall τ between the rankings for each of the 110 runs submitted to the track. If the τ and area-between-MAP-curves agreed as to which runs made good predictions, the points would lie on a line from the upper left to the lower right. While the general tendency is roughly in that direction, there are enough outliers to argue against using Kendall's τ over the entire topic ranking for this purpose.

Figure 2 also shows that there is quite a range in systems' abilities to predict which topics will be poor performers for them. Twenty-two of the 110 runs representing 5 of the 14 groups had area-between-MAP-curves scores of 0.5 or less. Thirty runs representing six groups (all distinct from the first group) had area-between-MAP-curves scores

[1] The run with the smallest difference was an ineffective run where almost all topics had very small average precision scores.



(a) run uic0404

(b) run humR04d5

Figure 1: Effect of differences in actual and predicted rankings on MAP scores.

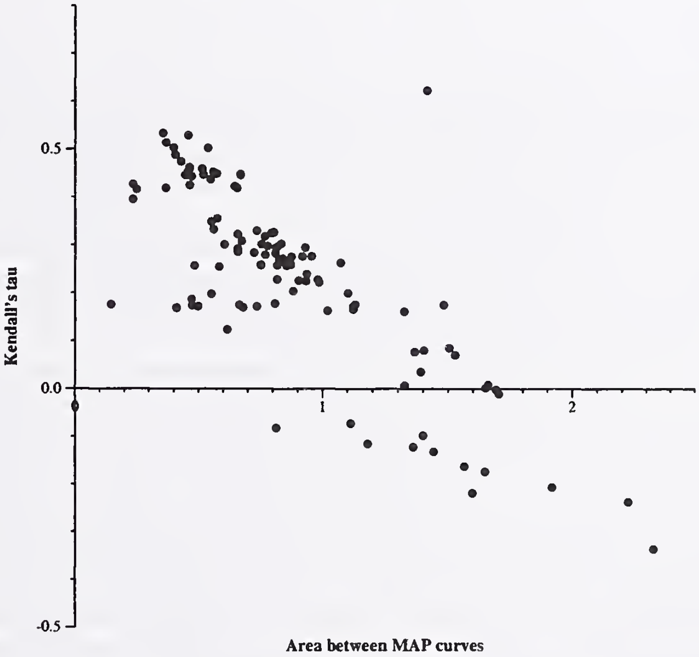


Figure 2: Scatter plot of area-between-MAP-curves vs. Kendall's τ for robust track runs.

of greater than 1.0 How much accuracy is required—including whether accurate predictions can be exploited at all—remains to be seen.

4 Evaluating Ineffectiveness

Most TREC topic sets contain 50 topics. In the TREC 2003 robust track we showed that the %no and area measures that emphasize poorly performing topics are unstable when used with topic sets as small as 50 topics. The problem is that the measures are defined over a subset of the topics in the set causing them to be much less stable than traditional measures for a given topic set size. In turn, the instability causes the margin of error associated with the measures to

Table 5: Error rate and proportion of ties for different measures and topic set sizes.

	50 Topics		75 Topics		100 Topics		124 Topics	
	Error Rate (%)	Proportion of Ties	Error Rate (%)	Proportion of Ties	Error Rate (%)	Proportion of Ties	Error Rate (%)	Proportion of Ties
MAP	2.4	0.144	1.3	0.146	0.7	0.146	0.3	0.145
P10	4.0	0.215	2.1	0.223	1.1	0.226	0.6	0.228
%no	14.1	0.107	11.8	0.146	9.6	0.064	7.6	0.065
area	10.6	0.040	7.9	0.041	5.9	0.042	4.7	0.042

be large relative to the difference in scores observed in practice.

4.1 Stability of %no and area measure

The motivation for using 250 topics in the this year's track was to test the stability of the measures on larger topic set sizes. The empirical procedures to compute the error rates and error margins are the same as were used in the 2003 track [11] except the topic set size is varied. Since the combined topic set contained 249 topics, topic set sizes up to 124 (half 249) can be tested.

Table 5 shows the error rate and proportion of ties computed for the four different measures used in table 3 and four different topic set sizes: 50, 75, 100, and 124. The error rate shows how likely it is that a single comparison of two systems using the given topic set size and evaluation measure will rank the systems in the wrong order. For example, an error rate of 3% says that in 3 out of 100 cases the comparison will be wrong. Larger error rates imply a less stable measure. The proportion of ties indicates how much discrimination power a measure has; a measure with a low error rate but a high proportion of ties has little power.

The error rates computed for topic set size 50 are somewhat higher than those computed for the TREC 2003 track, probably reflecting the greater variety of topics the error rate was computed from. The general trends in the error rates are strong and consistent: error rate decreases as topic set size increases, and the %no and area measures have a significantly higher error rate than MAP or P(10) at equal topic set sizes.

Using the standard of no larger than a 5% error rate, the area measure can be used with test sets of at least 124 topics, while the %no measure requires still larger topics sets. Note that since the area measure is defined using the worst quarter topics, a 124 topic set size implies the measure is using 31 topics in its computation. While this is good for stability, it is no longer as focused on the very poor topics.

The error rates shown in table 5 assumed two runs whose difference in score was less than 5% of the larger score were equally as effective. By using a larger value for the difference before deciding two runs are different, we can decrease the error rate for a given topic set size (because the discrimination power is reduced) [12]. Table 6 gives the critical value required to obtain no more than a 5% error rate for a given topic set size. For the area measure, the critical value is the minimum difference in area scores needed. For the %no measure, the critical value is the number of additional questions that must have no relevant in the top ten, also expressed as a percentage of the total topic set size. Also given in the table is the percentage of the comparisons that exceeded the critical value when comparing all pairs of runs submitted to the track over all 1000 topic sets used to estimate the error rates. This percentage demonstrates how sensitive the measure is to score differences encountered in practice.

The sensitivity of the %no measure does increase with topic set size, but the sensitivity is still very poor even at 124 topics. While intuitively appealing, this measure is just too coarse to be useful unless there are massive numbers of topics. Note that the same argument applies to the "Success@10" measure (i.e., the number of topics that retrieve a relevant document in the top 10 retrieved) that is being used to evaluate tasks such as home page finding and the document retrieval phase of question answering.

The sensitivity of the area measure is more reasonable. The area measure appears to be an acceptable measure for topic set sizes of at least 100 topics, though as mentioned above, its emphasis on the worst performing topics lessens as topic size grows.

Table 6: Sensitivity of measures: given is the critical value required to have an error rate no greater than 5% plus the percentage of comparisons over track run pairs that exceeded the critical value.

	50 Topics		75 Topics		100 Topics		124 Topics	
	Critical Value	% Significant	Critical Value	% Significant	Critical Value	% Significant	Critical Value	% Significant
%no	11 (22%)	3.8	16 (21%)	3.9	11 (10%)	15.7	13 (10%)	16.3
area	0.025	16.5	0.020	38.6	0.015	62.4	0.015	68.8

Table 7: Evaluation scores for the runs of Figure 3.

	MAP	geometric MAP	P10	area	%no
pircRB04td2	0.359	0.263	0.541	0.047	4
NLPR04clus10	0.306	0.230	0.449	0.048	8
uogRobLWR10	0.320	0.176	0.448	0.015	11

4.2 Geometric MAP

The problem with using MAP as a measure for poorly performing topics is that changes in the scores of better-performing topics mask changes in the scores of poorly performing topics. For example, the MAP of a run in which the effectiveness of topic A doubles from 0.02 to 0.04 while the effectiveness of topic B decreases 5% from 0.4 to 0.38 is identical to the baseline run's MAP. This suggests using a nonlinear rescaling of the individual topics' average precision scores before averaging over the topic set as a way of emphasizing the poorly performing topics.

The geometric mean of the individual topics' average precision scores has the desired effect of emphasizing scores close to 0.0 (the poor performers) while minimizing differences between larger scores. The geometric mean is equivalent to taking the log of the the individual topics' average precision scores, computing the arithmetic mean of the logs, and exponentiating back for the final geometric MAP score. Since the average precision score for a single topic can be 0.0—and trec_eval reports scores to 4 significant digits—we take the expedient of adding 0.00001 to all scores before taking the log (and then subtracting 0.00001 from the result after exponentiating).

To understand the effect of the various measures, Figure 3 shows a plot of the individual topic average precision scores for three runs from the TREC 2004 robust track. For each run, the average precision scores are sorted by increasing score and plotted in that order. Thus the x-axis in the figure represents a topic rank and the y-axis is the average precision score obtained by the topic at that rank. The three runs were selected to illustrate the differences in the measures. The pircRB04td2 run was the most effective run as measured by both standard MAP over all 249 topics and geometric MAP over all 249 topics. The NLPR04clus10 run has relatively few abysmal topics and also relatively few excellent topics, while the uogRobLWR10 run has relatively many of both abysmal and excellent topics. The evaluation scores for these three runs are given in Table 7. The uogRobLWR10 run has a better standard MAP score than the NLPR04clus10 run, and a worse area and geometric MAP score. The P(10) score for the two runs are essentially identical.

Table 8 shows that the geometric mean measure is also a stable measure. The table gives the error rate and proportion of ties for geometric MAP for various topic set sizes. As in Table 5, the geometric MAP's error rates are computed assuming a difference in scores less than 5% of the larger score is a tie. Compared to the error rates for the measures given in Table 5, geometric MAP's error rate is larger than both standard MAP and P(10) for equal topic set sizes, but much reduced compared to the area and %no measures. The geometric MAP measure has the additional benefit over the area measure of being less complex. Given just the geometric MAP scores for a run over two sets of topics, the geometric MAP score for that run on the combined set of topics can be computed, which is not the case for the area measure.

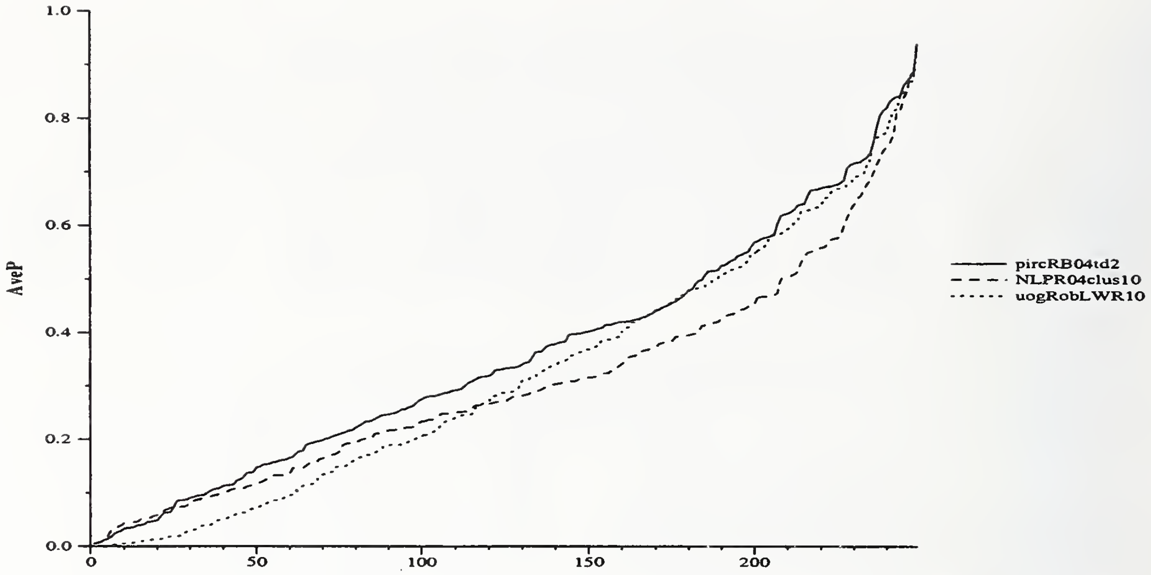


Figure 3: Individual topic average precision scores for three TREC 2004 runs.

Table 8: Error rate and proportion of ties computed over different topic set sizes for the geometric MAP measure.

Topic Set Size	Error Rate (%)	Proportion of Ties
25	9.1	0.081
50	5.2	0.086
63	4.1	0.088
75	3.4	0.090
100	2.3	0.092
124	1.5	0.094

5 Conclusion

The first two years of the TREC robust retrieval track have focused on trying to ensure that all topics obtain minimum effectiveness levels. The most promising approach to accomplishing this feat is exploiting text collections other than the target collection, usually the web. Believing that you cannot improve that which you cannot measure, the track has also examined evaluation measures that emphasize poorly performing topics. The geometric MAP measure is the most stable measure with a suitable emphasis.

The robust retrieval track is scheduled to run again in TREC 2005, though the focus of the track is expected to change. The current thinking is that the track will test the robustness of ad hoc retrieval technology by examining how stable it is in face of changes to the retrieval environment. To accomplish this, participants in the robust track will be asked to use their system for the ad hoc task in at least two of the other TREC tracks (for example, genomics and terabyte or terabyte and HARD). Within the robust track, same-system runs will be contrasted to see how differences in the tasks affect performance. Runs will also be evaluated using existing robust track measures, particularly geometric MAP.

Acknowledgements

Steve Robertson and Chris Buckley were instrumental in the development of the geometric MAP measure.

References

[1] Giambattista Amati, Claudio Carpineto, and Giovanni Romano. Fondazione Ugo Bordoni at TREC 2004. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

[2] Chris Buckley. Why current IR engines fail. In *Proceedings of the Twenty-Seventh Annual International ACM SIGIR Conference on Reserach and Development in Information Retrieval*, pages 584–585, 2004.

[3] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306, 2002.

[4] Donna Harman and Chris Buckley. The NRRC Reliable Information Access (RIA) Workshop. In *Proceedings of the Twenty-Seventh Annual International ACM SIGIR Conference on Reserach and Development in Information Retrieval*, pages 528–529, 2004.

[5] K.L. Kwok, L. Grunfeld, H.L. Sun, and P. Deng. TREC2004 robust track experiments using PIRCS. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

[6] Shuang Liu, Chaojing Sun, and Clement Yu. UIC at TREC-2004: Robust track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

[7] Christine Piatko, James Mayfield, Paul McNamee, and Scott Cost. JHU/APL at TREC 2004: Robust and terabyte tracks. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

[8] Vassilis Plachouras, Ben He, and Iadh Ounis. University of Glasgow at TREC2004: Experiments in web, robust and terabyte tracks with Terrier. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

[9] Alan Stuart. Kendall's tau. In Samuel Kotz and Norman L. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 4, pages 367–369. John Wiley & Sons, 1983.

[10] Ellen M. Voorhees. Overview of the TREC 2002 question answering track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, pages 57–68, 2003. NIST Special Publication 500-251.

[11] Ellen M. Voorhees. Overview of the TREC 2003 robust retrieval track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 69–77, 2004.

[12] Ellen M. Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–323, 2002.

[13] Jin Xu, Jun Zhao, and Bo Xu. NLPR at TREC 2004: Robust experiments. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

[14] Kiduk Yang, Ning Yu, Adam Wead, Gavin La Rowe, Yu-Hsiu Li, Christopher Friend, and Yoon Lee. WIDIT in TREC-2004 genomics, HARD, robust, and web tracks. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

[15] Elad Yom-Tov, Shai Fine, David Carmel, Adam Darlow, and Einat Amitay. Juru at TREC 2004: Experiments with prediction of query difficulty. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

Overview of the TREC 2004 Terabyte Track

Charles Clarke
University of Waterloo
claclark@plg.uwaterloo.ca

Nick Craswell
Microsoft Research
nickcr@microsoft.com

Ian Soboroff
NIST
ian.soboroff@nist.gov

Abstract

The Terabyte Track explores how adhoc retrieval and evaluation techniques can scale to terabyte-sized collections. For TREC 2004, our first year, 50 new adhoc topics were created and evaluated over a a 426GB collection of 25 million documents taken from the .gov Web domain. A total of 70 runs were submitted by 17 groups. Along with the top documents, each group reported average query times, indexing times, index sizes, and hardware and software characteristics for their systems.

1 Introduction

Early retrieval test collections were small, allowing relevance judgments to be based on an exhaustive examination of the documents but limiting the general applicability of the findings. Karen Sparck Jones and Keith van Rijsbergen proposed a way of building significantly larger test collections by using pooling, a procedure adopted and subsequently validated by TREC. Now, TREC-sized collections (several gigabytes of text and a few million documents) are small for some realistic tasks, but current pooling practices do not scale to substantially larger document sets. Thus, there is a need for an evaluation methodology that is appropriate for terabyte-scale document collections. A major research goal of the Terabyte track is to better define where our measures break down, and to explore new measures and methods for dealing with incomplete relevance judgments.

Current tasks that are evaluated using large web collections, such as known-item and high-precision searching, focus on the needs of the common web searcher but also arise from our inability to measure recall on very large collections. Good estimates of the total set of relevant documents are critical to the reliability and reusability of test collections as we now use them, but it would take hundreds of different systems, hundreds of relevance assessors, and years of effort to produce a terabyte-sized collection with completeness of judgments comparable to a typical TREC collection. Hence, new evaluation methodologies and ways of building test collections are needed to scale retrieval experiments to the next level.

The proposal for a TREC Terabyte Track was initiated at a SIGIR workshop in 2003 and accepted by the TREC program committee for TREC 2004. This report describes the details of the task undertaken, the runs submitted, and the range of approaches taken by the participants.

2 The Retrieval Task

The task is classic adhoc retrieval, a task which investigates the performance of systems searching a static set of documents using previously-unseen topics. This task is similar to the current Robust Retrieval task, and to the adhoc and VLC tasks from earlier TREC conferences.

2.1 Collection

This year's track used a collection of Web data crawled from Web sites in the .gov domain during early 2004. We believe that this collection ("GOV2") contains a large proportion of the crawlable pages in .gov, including HTML and text, plus the extracted text of PDF, Word and postscript files. By focusing the track on a single, large, interconnected domain we hoped to create a realistic setting, where content, structure and links could all be fruitfully exploited in the retrieval process.

The GOV2 collection is 426GB in size and contains 25 million documents. While this collection contains less than a full terabyte of data, it is considerably larger than the collections used in previous TREC tracks. For TREC 2004, the collection was distributed by CSIRO in Australia on a single hard drive for a cost of A$1200 (about US$800).

2.2 Topics

NIST created 50 new topics for the track. Figure 1 provides an example. As in the past, the title field may be treated as a keyword query, similar to the queries stereotypically entered by users of Web search systems. The description field provides a slightly longer statement of the topic requirements, usually expressed as a single complete sentence or question. Finally, the narrative supplies additional information necessary to fully specify the requirements, expressed in the form of a short paragraph. While keywords from the title are usually repeated in the description, they do not always appear in the narrative.

2.3 Queries

For each topic, participants created a query and submitted a ranking of the top 10,000 documents for that topic. Queries could be created automatically or manually from the topic statements. As for all TREC tasks, automatic methods are those in which there is no human intervention at any stage, and manual methods are everything else. For most runs, groups could use any or all of the topic fields when creating queries from the topic statements. However, each group submitting an automatic run was required to submit an automatic run that used just the title field.

2.4 Submissions

Each group was permitted to submit up to five experimental runs. Each run consists of the top 10,000 documents for each topic, along with associated performance and system information. We required 10,000 documents, since we believe this that information may useful during later analysis to help us better understand the evaluation process.

In addition to the top 10,000 documents, we required each group to report details of their hardware configuration and various performance numbers, including the number of processors, total RAM (GB), on-disk index size (GB), indexing time (elapsed time in minutes), average search time (seconds), and hardware cost. For the number of processors, we requested the total number of CPUs in the system, regardless of their location. For example, if a system is a cluster of eight

```
<top>
<num> Number: 705

<title>
Iraq foreign debt reduction

<desc> Description:
Identify any efforts, proposed or undertaken, by world governments to seek
reduction of Iraq's foreign debt.

<narr> Narrative: Documents noting this subject as a topic for
discussion (e.g. at U.N. and G7) are relevant. Money pledged for
reconstruction is irrelevant.

</top>
```

Figure 1: Terabyte Track Topic 705

dual-processor machines, the number of processors is 16. For the hardware cost, we requested an estimate in US dollars of the cost at the time of purchase.

Some groups may subset a collection before indexing, removing selected pages or portions of pages to reduce its size. Since subsetting may have an impact on indexing time and average query time, we asked each group to report the fraction of pages indexed.

For search time, we asked the groups to report the time to return the top 20 documents, not the time to return the top 10,000, since this number better reflects the performance that would be seen by a user. It was acceptable to execute a system twice for each query, once to generate the top 10,000 documents and once to measure the execution time for the top 20, provided that the top 20 results were the same in both cases.

2.5 Judgments

The top 85 documents of two runs from each group were pooled and judged by NIST assessors. The judgments used a three-way scale of "not relevant", "relevant", and "highly relevant".

3 Submitted Runs

Figures 2 and 3 provide an overview submitted runs. The first two columns give the group and run ids. The third column lists the topic fields — Title ("T"), Description ("D") and Narrative ("N") — that were used to create the query. In all cases queries were generated automatically from these fields. No manual runs were submitted. The next three columns indicate if link analysis techniques, anchor text, or other document structure was used in the ranking process. The third-last column gives the average query time required to generate the top 20 results, and the second-last column gives the time to build the index in hours. The last column gives the mean average precision achieved by each run.

Group Id	Run Id	Topic Fields	Links?	Anchors?	Structure?	Query Time (s)	Index Time (h)	MAP
cmu.dir.callan	cmuapfs2500	TDN	N	N	N	600	20.0	0.284
	cmutufs2500	T	N	N	N	240	20.0	0.248
	cmutuns2500	T	N	N	N	75	20.0	0.207
dubblincity.u	DcuTB04Base	T	N	N	N	2	408.7	0.118
	DcuTB04Ucd1	TDN	N	Y	N	84	883.7	0.076
	DcuTB04Wbm25	T	N	N	Y	2	760.8	0.079
	DcuTB04Combo	T	N	Y	Y	2	906.0	0.033
	DcuTB04Ucd2	TDN	N	Y	N	15	457.5	0.070
etymon	nn04tint	T	N	N	N	25	44.8	0.112
	nn04eint	T	N	N	N	78	44.8	0.074
	nn04test	T	N	N	N	46	44.8	0.028
hummingbird	humT04l	T	N	N	Y	115	100.0	0.224
	humT04dvl	T	N	N	Y	142	100.0	0.212
	humT04vl	T	N	N	Y	119	100.0	0.221
	humT04l3	T	N	N	Y	49	100.0	0.155
	humT04	T	N	N	Y	50	100.0	0.196
iit	iit00t	T	N	N	N	23	8.0	0.210
	robertson	T	N	N	N	42	8.0	0.200
jhu.apl.mcnamee	apl04w4tdn	TDN	N	N	N	10000	0.0	0.034
	apl04w4t	T	N	N	N	10000	0.0	0.027
max-planck.theobald	mpi04tb07	T	Y	N	Y	6	42.0	0.125
	mpi04tb09	TD	Y	N	Y	9	42.0	0.123
	mpi04tb101	TD	Y	N	N	9	42.0	0.081
	mpi04tb81	TD	Y	N	N	9	42.0	0.092
	mpi04tb91	TD	Y	N	N	9	42.0	0.092
microsoft.asia	MSRAt3	T	N	Y	Y	1	11.6	0.171
	MSRAt4	T	N	Y	Y	1	11.6	0.188
	MSRAt5	T	N	Y	Y	1	11.6	0.190
	MSRAt2	T	N	N	Y	1	11.6	0.092
	MSRAt1	T	N	N	Y	1	11.6	0.191
rmit.scholer	zetbodoffff	T	N	N	N	25	13.5	0.219
	zetanch	T	N	Y	N	2	13.6	0.217
	zetplain	T	N	N	N	2	13.5	0.223
	zetfuzzy	T	N	Y	N	2	13.6	0.131
	zetfunkyz	T	N	Y	N	3	13.6	0.207

Figure 2: Summary of Submitted Runs (Part 1)

Group Id	Run Id	Topic Fields	Links?	Anchors?	Structure?	Query Time (s)	Index Time (h)	MAP
sabir.buckley	sabir04td3	D	N	N	N	18	14.0	0.117
	sabir04ta2	TDN	N	N	N	9	14.0	0.172
	sabir04tt	T	N	N	N	1	14.0	0.116
	sabir04td2	D	N	N	N	3	14.0	0.121
	sabir04tt2	T	N	N	N	1	14.0	0.118
tsinghua.ma	THUIRtb5	T	N	N	N	15	32.0	0.244
	THUIRtb4	TDN	N	Y	N	55	17.0	0.245
	THUIRtb3	T	N	Y	N	9	17.0	0.220
	THUIRtb2	TDN	N	Y	Y	18	2.8	0.056
	THUIRtb6	T	N	N	N	16	32.0	0.204
u.alaska	irttbtl	T	N	N	Y	5	30.0	0.009
u.amsterdam.lit	UAmsT04TBm1	T	N	Y	Y	90	4.3	0.044
	UAmsT04TBanc	T	N	Y	N	1	0.3	0.013
	UAmsT04TBm1p	T	N	Y	Y	90	4.3	0.043
	UAmsT04TBtit	T	N	N	Y	20	4.0	0.039
	UAmsT04TBm3	T	N	Y	Y	90	4.3	0.043
u.glasgow	uogTBQEL	TDN	N	N	N	46	200.6	0.307
	uogTBPoolQEL	TDN	N	N	N	46	200.6	0.231
	uogTBBaseS	T	N	N	N	4	200.6	0.271
	uogTBAnchS	T	N	Y	N	3	501.7	0.269
	uogTBBaseL	TDN	N	N	N	28	200.6	0.305
u.mass	indri04AWRM	T	N	N	N	39	5.9	0.284
	indri04AW	T	N	N	N	7	5.9	0.269
	indri04QLRM	T	N	N	N	26	5.9	0.253
	indri04QL	T	N	N	N	1	5.9	0.251
	indri04FAW	T	N	Y	Y	52	21.6	0.279
u.melbourne	MU04tb3	T	Y	Y	N	0.08	2.5	0.043
	MU04tb2	T	N	Y	N	0.08	2.5	0.063
	MU04tb4	T	Y	Y	N	0.36	13.0	0.268
	MU04tb1	T	N	N	N	0.08	1.7	0.266
	MU04tb5	T	Y	Y	N	0.08	2.5	0.064
upisa.attardi	pisa4	T	Y	Y	Y	3	16.0	0.103
	pisa3	T	Y	Y	Y	3	16.0	0.107
	pisa2	T	Y	Y	Y	3	16.0	0.096
	pisa1	T	Y	Y	Y	1	16.0	0.050

Figure 3: Summary of Submitted Runs (Part 2)

4 Overview of Systems

Most groups contributed papers to this notebook, and we refer the reader to the these papers for complete details about individual systems. In .the remainder of this section, we summarize the range of approaches taken by the groups and highlight some unusual features of their systems.

4.1 Hardware and Software

The cost and scale of the hardware varied widely, with many groups dividing the documents across multiple machines and searching the collection in parallel. At one extreme, the group from the University of Alaska's Arctic Region Supercomputing Center used 40 nodes of the NCSA "mercury" TeraGrid cluster, which cost over US$10 million. At the other extreme, the group from Tsinghua University used a single PC with an estimated cost of US$750.

To index and search the collection, most groups used custom retrieval software develop by their own group or by an associated group. One exception is the University of Alaska, which used MySQL (finding a bug in the process). Hummingbird used their commercial SearchServer^{tm} system. Etymon Systems used their Amberfish package, which they have released as open source (`etymon.com/tr.html`). Both CMU and University of Massachusetts used Indri, a new indexing and retrieval component developed by the University of Massachusetts for the Lemur Toolkit.

4.2 Indexing

Overall, indexing methods were fairly standard. Most groups applied stopping and stemming methods. However, at least three groups, the University of Massachusetts, CMU, and Etymon Systems did not remove stopwords, despite the size of the collection. Several groups compressed the index to improve performance and reduce storage requirements, including the University of Glasgow, the University of Melbourne, and the University of Pisa. Sabir implemented compressed indices, but did not use them in their final runs.

Since a large portion the collection consists of HTML, many groups applied special processing to the anchor text or to specific fields within the documents. For example, Dublin City University generated surrogate anchor text documents, comprised of the anchor text of inlinks to a document. The Indri system supports the indexing of arbitrary document fields, and this facility was used to index various fields of HTML documents (title, h1, h2, etc.). The University of Pisa performed extensive preprocessing, extracting page descriptions and categories from Dmoz, collecting links and anchor texts, and identifying specific fields within HTML documents.

The most unusual approach was taken by the University of Amsterdam group, who indexed only document titles and anchor text. The resulting indexes are small: 1.4GB for the titles covering 83% of the documents, and 0.1 GB for the anchors covering 6% of the documents. This very selective indexing produced a 20 minute indexing time and a 1 second average query time without the need for special performance optimizations.

Figure 4 plots the fastest indexing times, ignoring all but the fastest time from each group. Indexing a 426GB collection in under 14 hours implies an indexing rate of over 30GB/hour. However, most of these groups parallelized the indexing process or indexed only a subset of the collection. The fastest reported "indexing" time, zero, does not appear on the figure. The group reporting this indexing time, JHU/APL, did not index the collection at all. Instead, they searched it with a DFA executed by a Perl script.

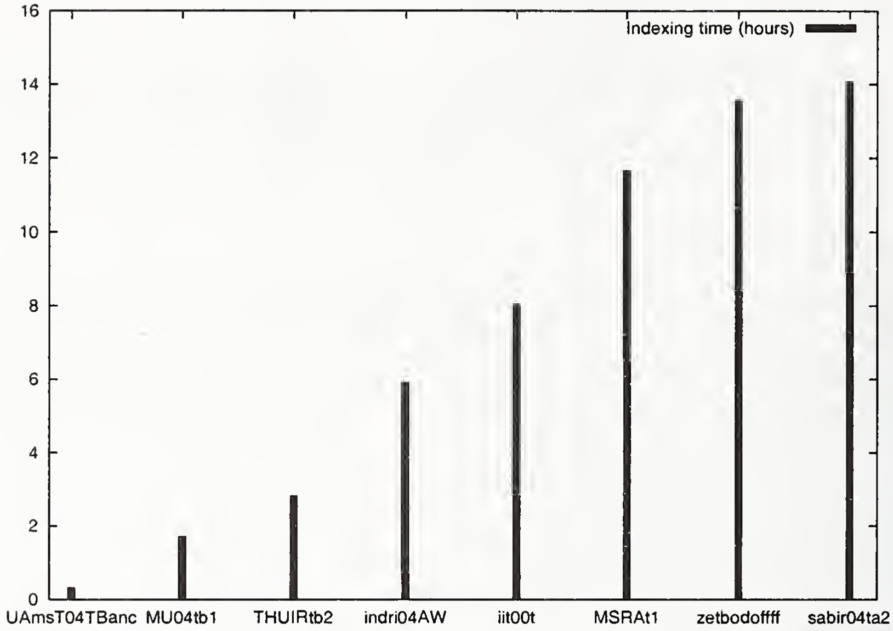


Figure 4: Indexing Time (hours) — Top 8 Groups

4.3 Query Processing

Although adhoc retrieval has been a mature technology for many years, a surprising variety of retrieval formulae were used, including Okapi BM25, cosine, and methods based on language modeling and divergence from randomness. Proximity operators were used by several groups including University of Pisa and CMU. Link analysis methods were used in 17% of the runs, anchor text was used in 37%, and other document structure (usually document titles) was used in 36%. Several groups expanded queries using pseudo-relevance feedback. This wide range of methods suggests that "best practice" for information retrieval over large Web collections may not be as well established as some believe.

Figure 5 plots the eight fastest average query times, ignoring all but the fastest run from each group. The run submission form requested the average query time in seconds, rather than milliseconds, and the impact of this error can be seen in the figure. Five groups reported an average query time of "1 second" and two groups reported a time of "2 seconds". The query time reported by the University of Melbourne, 0.08 seconds, is roughly equal to the time typically required for a single disk access.

Figure 6 plots the title-only runs achieving the best mean average precision, ignoring all but the best-performing run from each group. The curve is relatively flat, with all eight groups achieving reasonable performance.

5 The Future

For TREC 2005, the Terabyte Track will continue to use the GOV2 collection, giving us a total of 100 topics over the collection. We plan to collect more and better information regarding system performance, with the hope that system performance comparisons can be made more realistically. Finally, a known-item retrieval task may be added to the track.

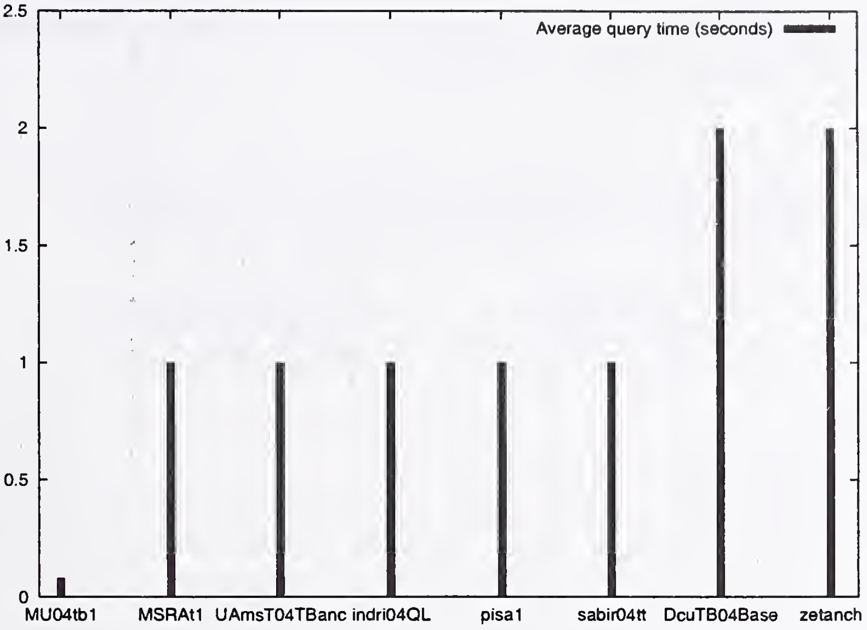


Figure 5: Average Query Time (seconds) — Top 8 Groups

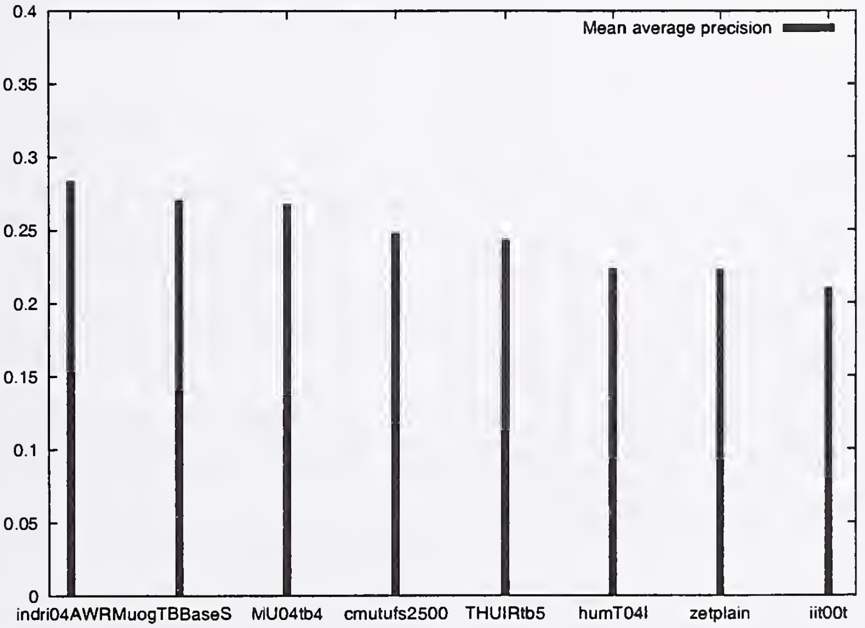


Figure 6: Mean Average Precision (MAP) — Top 8 Groups

6 Acknowledgments

We thank Doug Oard for hosting our Web crawler at the University of Maryland, and CSIRO for processing and distributing the GOV2 collection.

Overview of the TREC-2004 Web Track

Nick Craswell
MSR Cambridge, UK
nickcr@microsoft.com

David Hawking
CSIRO, Australia
david.hawking@csiro.au

1 Introduction

This year's main experiment involved processing a mixed query stream, with an even mix of each query type studied in TREC-2003: 75 homepage finding queries, 75 named page finding queries and 75 topic distillation queries. The goal was to find ranking approaches which work well over the 225 queries, without access to query type labels.

We also ran two small experiments. First, participants were invited to submit classification runs, attempting to correctly label the 225 queries by type. Second, we invited participants to download the new W3C test collection, and think about appropriate experiments for the proposed TREC-2005 Enterprise Track. This is the last year for the Web Track in its current form, it will not run in TREC-2005.

2 Mixed query task

The mixed query task was conducted using the 18 gigabyte, 1.25 million document crawl of the .GOV domain. Last year's tasks involved queries of three types:

Topic distillation The query describes a general topic, e.g. 'electoral college', the system should return homepages of relevant sites.

Homepage finding The query is the name of a site that the user wishes to reach, e.g. 'Togo embassy', and the system should return the URL of that site's homepage at (or near) rank one.

Named page finding The query is the name of a non-homepage that the user wishes to reach, e.g. 'Ireland consular information sheet', and the system should return the URL of that page at (or near) rank one.

There are several possible approaches to dealing with the mixed query stream. One is to find a robust ranking method which works well for all three types. Another is to find specialised methods e.g. one for TD, one for NP and one for HP. Specialised methods could be combined, for example by interleaving ranks or combining scores. Combination can either be done uniformly for all queries or based on query classification, preferring the specialist method which seems most appropriate for the current query.

2.1 Judging and Measures

Since each NP and HP topic is developed with a URL in mind, the only judging task is to identify URLs of equivalent (near-duplicate) pages. For example identifying that `http://xyz.gov/` and `http://xyz.gov/index.html` are equivalent answers. TD judging is more time consuming. Finding URLs which are homepages of relevant sites involves a relevance judgment combined with understanding of site structure, which can be gained by navigating between pages and looking at URL(s).

Judges found 1763 relevant[1] pages: 80 for NP (5 extra), 83 for HP (8 extra) and 1600 for TD. For distillation, the mean number of results per query was $1600/75 = 21.3$, with a median of 13. Topic distillation 2003 had mean 10.3 and median 8. Because there were no major changes in query development and judging methods, we believe the 2003 and 2004 sets are matching and reusable test sets for topic distillation.

We have four measures which we can apply to all query types:

MAP and MRR Mean average precision (MAP) and

[1] Varying the definition of relevant according to the query type.

mean reciprocal rank of the first correct answer (MRR) are standard TREC measures. They are related measures, in that they are exactly equivalent for queries with one correct answer. The problem with applying MAP globally is that some NPHP queries have multiple answers and we only care about the first correct answer. Therefore we apply MAP to TD queries and MRR to NPHP queries. Both measures are calculated on the whole run (1000 ranks), but both put a natural emphasis on the top-ranked documents.

Success@1 The proportion of queries for which a good answer was at rank 1 (the first result the user sees).

Success@5 The proportion of queries for which one or more good answers were in the top 5. The top 5 is what might typically appear on the results page of a web search system, without the user needing to scroll ("above the fold"). If a correct answer appears in the top 5 for 90 of 225 queries, then S@5=0.4.

Success@10 This measure indicates how often a system found something in the top 10, which typically is the first page of web search results. This can also be thought of as a failure measure, because $1 - \text{S@10}$ is the proportion of queries with nothing in the top 10.

We also apply Precision@10 and Recall@1000 to the topic distillation queries.

2.2 Results per query type

Table 3 presents the results for the 75 distillation queries. Considering the MAP and P@10 measures, the top two groups tied, only differing by 0.0011 in MAP and 0.0014 in P@10. Groups 3 and 4 are also very close to each other.

Table 1 has the results for the 75 named page queries. This year's NP MRR scores are higher than last year's, but a striking difference is that the gap between NP and HP has closed. This is illustrated in Figure 1 which, compared to a similar plot last year, has a much smaller gap between HP and NP for the top-scoring runs. This could reflect a better balance between 'relevance' and homepage bias (too much homepage bias hurts NP performance).

Table 2 shows results for HP queries. Although the results are high, they are not as high as last year's best HP

Run	MRR	S@1	S@5	S@10
MSRC04B2S	0.731	0.653	0.827	0.880
MSRAx4	0.685	0.587	0.787	0.853
UAmsT04MSind	0.640	0.507	0.800	0.867
uogWebSelAnL	0.619	0.493	0.787	0.840
THUIRmix045	0.619	0.493	0.787	0.867
MeijiHILw1	0.611	0.480	0.800	0.867
ICT04CIIS1AT	0.606	0.480	0.760	0.880
humW04pl	0.569	0.480	0.667	0.760
wdf3oks0a	0.545	0.413	0.693	0.760
SJTUINCMIX2	0.543	0.387	0.733	0.787
VTOK5	0.511	0.400	0.640	0.733
csiroatnist	0.456	0.320	0.613	0.680
mpi04web08	0.423	0.347	0.507	0.547
MU04web5	0.411	0.333	0.493	0.560
LamMcm1	0.323	0.213	0.440	0.547
fdwiedf0	0.276	0.147	0.453	0.533
irtbow	0.159	0.120	0.173	0.293
XLDBTumba01	0.068	0.067	0.067	0.080

Table 1: Named page results.

Run	MRR	S@1	S@5	S@10
MSRC04C12	0.749	0.653	0.840	0.880
MSRAx2	0.729	0.653	0.867	0.907
UAmsT04MSinu	0.659	0.560	0.760	0.827
THUIRmix045	0.626	0.533	0.733	0.787
uogWebSelAnL	0.625	0.493	0.813	0.840
csiroatnist	0.568	0.467	0.680	0.747
ICT04MNZ3	0.563	0.467	0.653	0.747
MU04web1	0.553	0.467	0.667	0.693
SJTUINCMIX3	0.489	0.400	0.613	0.667
humW04rdpl	0.479	0.373	0.587	0.693
MeijiHILw1	0.473	0.360	0.640	0.680
wdf3oks0brr1	0.421	0.320	0.493	0.640
mpi04web08	0.379	0.307	0.467	0.493
fdwiedf0	0.379	0.333	0.413	0.493
LamMcm1	0.326	0.267	0.413	0.453
VTOK5	0.270	0.173	0.373	0.427
irttil	0.090	0.053	0.120	0.173
XLDBTumba01	0.004	0.000	0.013	0.013

Table 2: Homepage results.

Run	MAP	P@10	R@1000	S@1	S@5	S@10
uogWebCAU150	0.179	0.249	0.777	0.507	0.773	0.893
MSRAmixed1	0.178	0.251	0.815	0.387	0.720	0.880
MSRC04C12	0.165	0.231	0.744	0.387	0.747	0.800
humW04rdpl	0.163	0.231	0.808	0.373	0.787	0.907
THUIRmix042	0.147	0.205	0.761	0.213	0.587	0.747
UAmsT04MWScb	0.146	0.209	0.786	0.360	0.667	0.760
ICT04CIIS1AT	0.141	0.208	0.785	0.333	0.640	0.787
SJTUINCMIX5	0.129	0.189	0.748	0.293	0.573	0.720
MU04web1	0.115	0.199	0.647	0.333	0.640	0.760
MeijiHILw3	0.115	0.153	0.547	0.307	0.547	0.640
csiroatnist	0.111	0.205	0.261	0.320	0.693	0.853
mpi04web01	0.106	0.177	0.453	0.240	0.640	0.787
VTOK5	0.101	0.135	0.721	0.187	0.493	0.533
fdwiedf0	0.090	0.117	0.536	0.293	0.493	0.587
wdf3oks0brr1	0.085	0.124	0.720	0.120	0.413	0.573
LamMcm1	0.049	0.087	0.270	0.173	0.400	0.467
irttil	0.018	0.029	0.147	0.067	0.147	0.173
XLDBTumba01	0.003	0.011	0.008	0.040	0.093	0.107

Table 3: Distillation results.

Run	Average	TD MAP	NP MRR	HP MRR	S@1	S@5	S@10
MSRC04B2S	0.546	0.162	0.731	0.745	0.564	0.809	0.862
MSRAx4	0.527	0.175	0.685	0.721	0.516	0.796	0.871
UAmsT04MSind	0.477	0.133	0.640	0.657	0.453	0.733	0.818
uogWebSelAn	0.466	0.166	0.615	0.617	0.444	0.760	0.818
THUIRmix045	0.457	0.126	0.619	0.626	0.409	0.702	0.778
ICT04MNZ3	0.435	0.137	0.603	0.563	0.440	0.689	0.769
MeijiHILw1	0.398	0.110	0.611	0.473	0.364	0.671	0.738
SJTUINCMIX2	0.385	0.125	0.543	0.487	0.347	0.618	0.689
csiroatnist	0.378	0.111	0.456	0.568	0.369	0.662	0.760
humW04rdpl	0.375	0.163	0.484	0.479	0.369	0.671	0.782
wdf3oks0arr1	0.344	0.085	0.542	0.404	0.276	0.542	0.653
MU04web1	0.343	0.115	0.362	0.553	0.356	0.587	0.662
mpi04web08	0.295	0.082	0.423	0.379	0.298	0.520	0.564
VTOK5	0.294	0.101	0.511	0.270	0.253	0.502	0.564
fdwiedf0	0.248	0.090	0.276	0.379	0.258	0.453	0.538
LamMcm1	0.232	0.049	0.323	0.326	0.218	0.418	0.489
irtbow	0.086	0.012	0.159	0.086	0.071	0.133	0.231
XLDBTumba01	0.025	0.003	0.068	0.004	0.036	0.058	0.067

Table 4: Overall results. Average is the mean of the TD MAP, NP MRR and HP MRR.

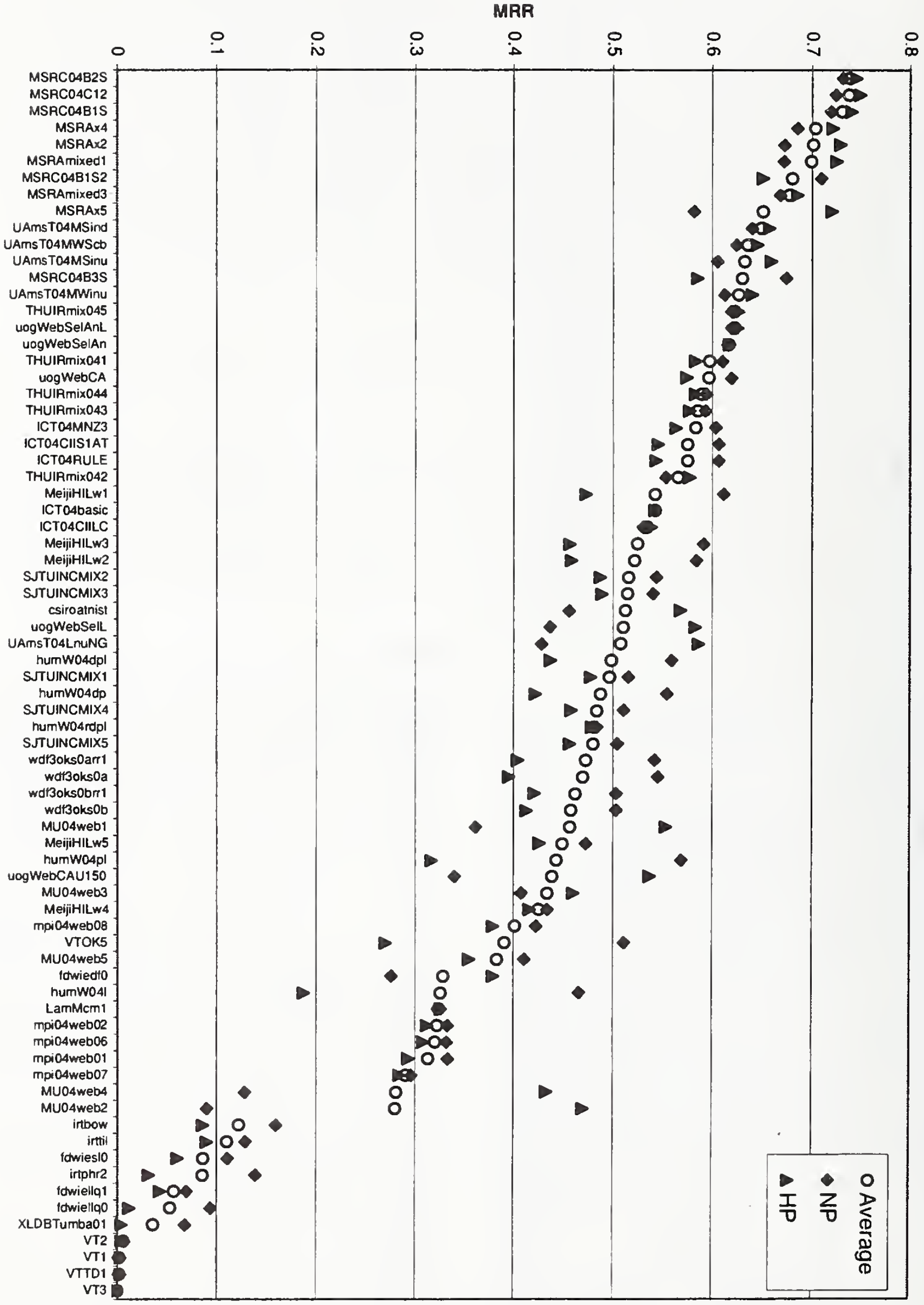


Figure 1: This year the top runs had less of a gap between HP and NP performance (compared to a plot in last year's overview).

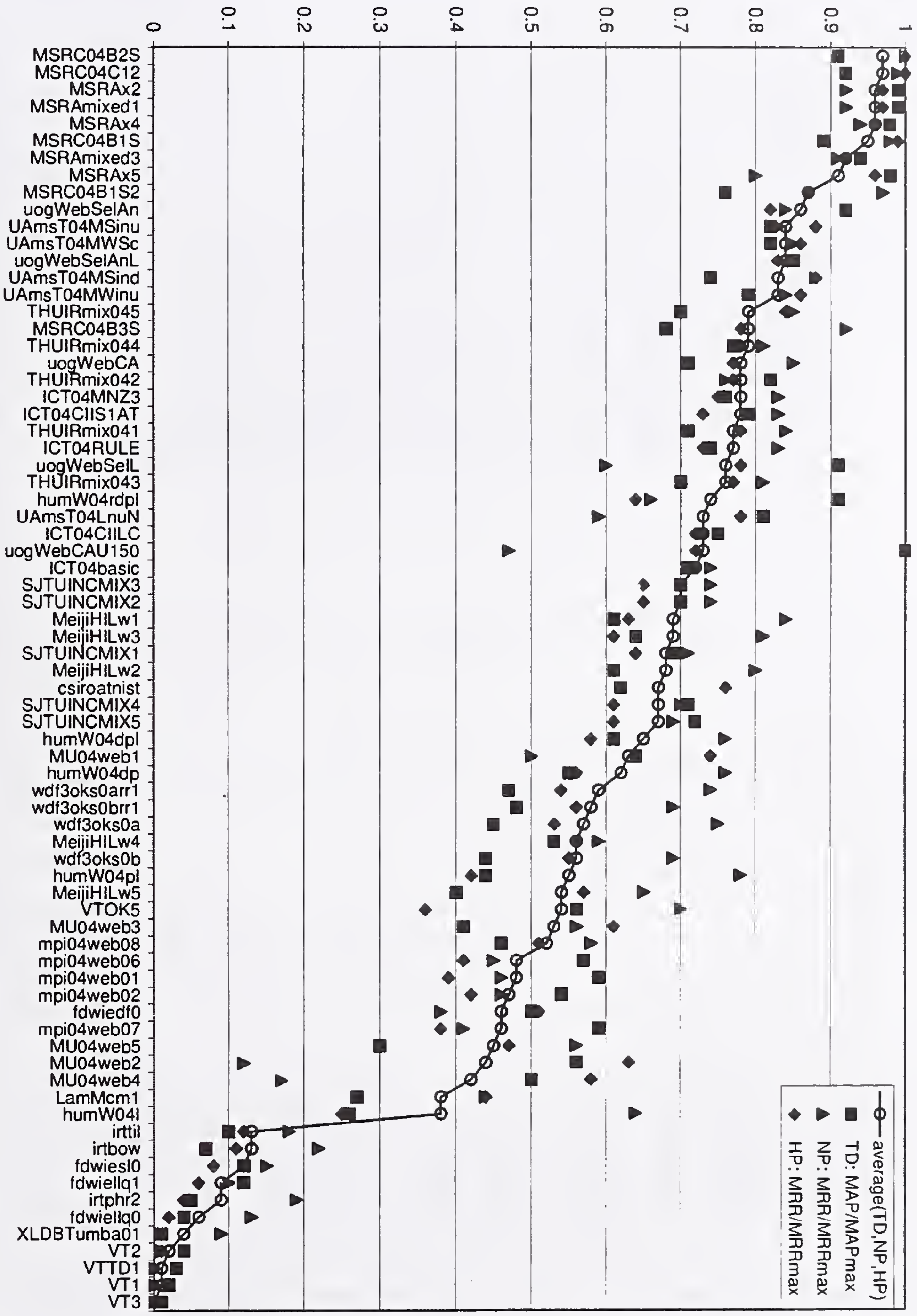


Figure 2: Performance of all runs, based on ratios with the best run of each type.

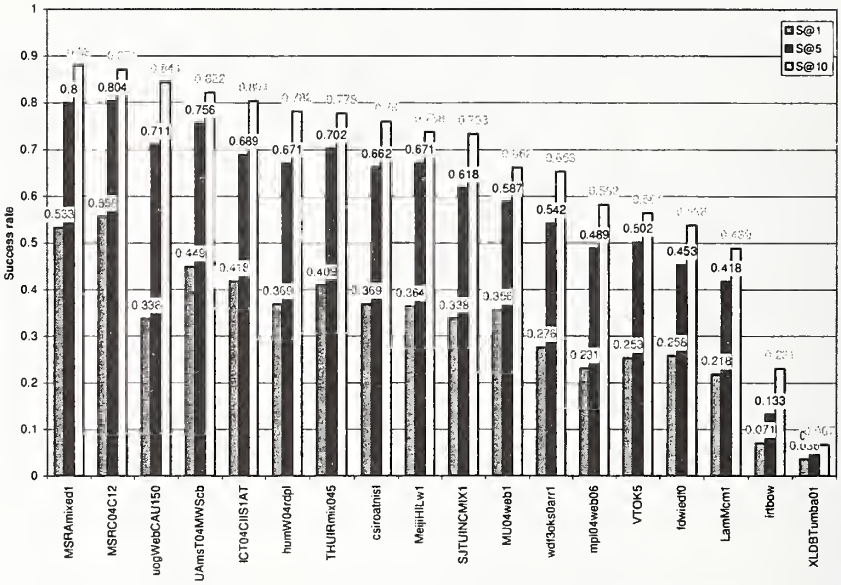


Figure 3: Success rate results. Best run from each group, by S@10.

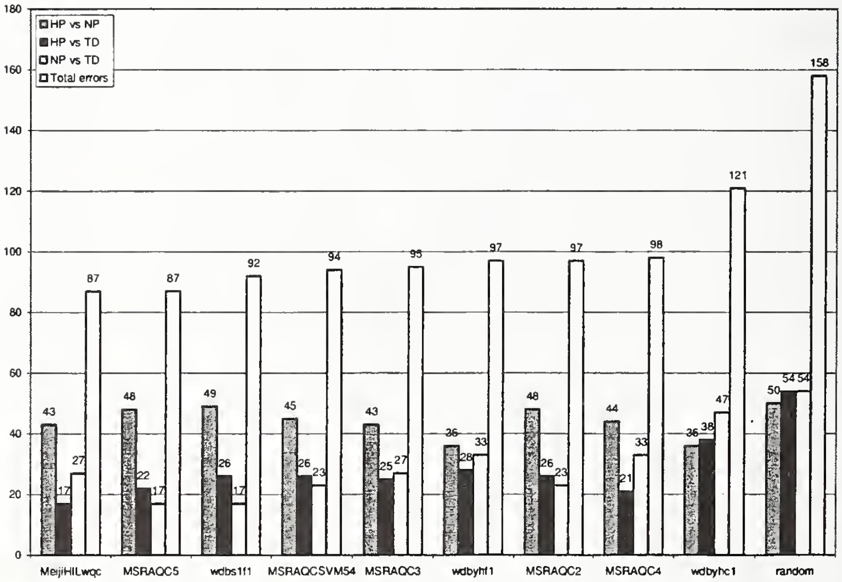


Figure 4: Results of query classification runs. Three types of error and total error.

Run	Avg	TD MAP	NP MRR	HP MRR	Anc	Lnk	Strc	ULen	UOth	QCls
MSRC04C12	0.97	0.92 (0.165)	0.99 (0.724)	1.00 (0.749)	yes	yes	yes	yes	no	no
MSRAx2	0.96	0.99 (0.177)	0.92 (0.672)	0.97 (0.729)	yes	yes	yes	yes	yes	no
uogWebSelAn	0.86	0.92 (0.166)	0.84 (0.615)	0.82 (0.617)	yes	no	yes	yes	no	yes
UAmsT04MWScb	0.84	0.82 (0.146)	0.85 (0.624)	0.86 (0.645)	yes	yes	yes	yes	no	no
THUIRmix045	0.79	0.70 (0.126)	0.85 (0.619)	0.84 (0.626)	yes	no	yes	no	no	no
ICT04CIIS1AT	0.78	0.79 (0.141)	0.83 (0.606)	0.73 (0.545)	yes	no	yes	no	no	no
humW04rdpl	0.74	0.91 (0.163)	0.66 (0.484)	0.64 (0.479)	no	no	yes	yes	yes	no
SJTUINCMIX3	0.70	0.70 (0.125)	0.74 (0.540)	0.65 (0.489)	yes	no	yes	no	no	yes
MeijiHILw1	0.69	0.61 (0.110)	0.84 (0.611)	0.63 (0.473)	yes	yes	yes	yes	no	no
csiroatnist	0.67	0.62 (0.111)	0.62 (0.456)	0.76 (0.568)	yes	yes	yes	yes	yes	no
MU04web1	0.63	0.64 (0.115)	0.50 (0.362)	0.74 (0.553)	yes	yes	yes	yes	yes	no
wdf3oks0arr1	0.59	0.47 (0.085)	0.74 (0.542)	0.54 (0.404)	yes	no	yes	yes	yes	no
VTOK5	0.54	0.56 (0.101)	0.70 (0.511)	0.36 (0.270)	yes	no	yes	no	yes	no
mpi04web08	0.52	0.46 (0.082)	0.58 (0.423)	0.51 (0.379)	yes	yes	yes	yes	yes	no
fdwiedf0	0.46	0.50 (0.090)	0.38 (0.276)	0.51 (0.379)	no	no	no	yes	yes	no
LamMcm1	0.38	0.27 (0.049)	0.44 (0.323)	0.44 (0.326)	yes	yes	yes	yes	yes	no
irtbow	0.13	0.07 (0.012)	0.22 (0.159)	0.11 (0.086)	no	no	no	no	no	no
XLDBTumba01	0.04	0.01 (0.003)	0.09 (0.068)	0.01 (0.004)						

Table 5: Normalised overall results with indication of methods used. Anc: Anchor text used? Lnk: Other link structure used? Strc: Document structure used? ULen: URL length used? UOth: Other URL features used? QCls: Special processing for different query types?

performance, of nearly 0.80. Similarly to last year, S@10 performance seems to max out at around 90%.

2.3 Overall results

Table 4 presents the best run from each group, judged on the average of TD MAP, NP MRR and HP MRR. Although the magnitude for TD is much less than NP and HP, MAP and MRR are related measures so it makes sense to look at the average.

Another way to get an overall score out of TD MAP, NP MRR and HP MRR is to normalise each query type according to the maximum score. This gives each run three scores between 0 and 1, and the average of these three scores is an overall score. Such scores are presented in Figure 2 and Table 5.

A third way to look at the overall result is by success rate. Success at 10 is an interesting number, because it is different from MAP and MRR which give a lot of weight to rank one, and it indicates how often a user reads a whole page of results without finding a good answer. Figure 3 presents success rate figures for the best run from each group, according to S@10 across all queries. The best S@10=0.88 measure gives the user no useful documents for 12% of queries, although perhaps this is acceptable if we assume that in those cases the user reformulates their query.

2.4 What worked

Table 5 indicates which technologies were used by the best run from each group. It is clear that most groups use document structure and many use anchor text. It also seems useful to use link structure and URL length. Other URL features and query classification were not necessary for good performance, but if groups had their best run using such methods they may well be helpful.

We also present information on methods used by the best run from several groups. (Full information is in Appendix A.)

1. **MSRC04C12** Interleaving of stem and nostem runs, each using structure, URL length and PageRank.

3. **MSRAx2** We interpolated relevance scores on the fields of title, body, anchor, url and merged the former four together. The score functions include BM25, proximity and a new proposed URL score function. And the final score combines relevance score and a HostRank that is a PageRank-like value.

10. **uogWebSelAn** Content and anchor-text retrieval, Porter Stemming, Divergence From Randomness PL2 weighting scheme, URL-length reranking, Selecting between content and anchor-text retrieval, or content with anchor-text and URL-length reranking

11. **UAmsT04MWScb** CombMNZ (non-normalized, non-weighted) of stemmed and non-stemmed runs, each using a mixture language model on stemmed full-text, titles, and anchor texts, using both an indegree and URL prior.

16. **THUIRmix045** Word pair weighting based on another run, which used content retrieval in full text and in-link anchor, with a larger weight in fields of Title, head, Bold and first line of page content.

20. **ICT04CIIS1AT** Anchor text forward propagation, page title text back propagation, combination of anchor text ,key words ,h1 text etc. ,different pivoted weigth function for different part

27. **humW04rdpl** Plain content search including linguistic expansion from English inflectional stemming, extra weight on properties such as Title and Metadata, lower url depth and root urls

3 Query classification runs

Three groups submitted a total of 9 query classification runs. Results are presented in Figure 4. Random classification of 225 queries into three types would tend to lead to about 150 errors, so classification runs were able to do significantly better than random. The best run MeijiHILwqc was a manual run. The most common type of error was confusing HP and NP (either by classifying HP as NP or classifying NP as HP).

4 W3C Investigation

Workshop participants proposed a variety of new experiments, for example relevance ranking in email, or searching for people who are experts in a particular topic area. We plan to pursue such ideas using the W3C dataset in the TREC-2005 Enterprise Track.

5 Conclusion

The main experiment showed that, on a mixed query set, effective retrieval is possible without query classification. Topic distillation is still by far the most difficult query type. Query classification runs showed that it is indeed possible to tell the difference. The most common classification mistake was to confuse NP and HP queries.

The other effect of the mixed query task is to consolidate the findings of previous Web Track years. There are web search information needs which are based on a page's position (a 'homepage') and importance, rather than just the page's text. To answer these information needs, it is not sufficient to search on content alone: use of 'Web evidence' based on structure, links and URLs is necessary. This evidence may be effectively used in an enterprise-scale crawl, of a million pages. The Web Track collections are now reusable resources for new experiments with TD, NP, HP and mixed query streams.

Of course there is also more work to be done in developing evaluation methodologies. Future web experiments could model other user needs, for example transactional search, and refine solutions to tricky issues such as distillation judging and scoring of near-duplicate results. Another direction would be to venture into the wider Web, where adversarial information retrieval is an issue, and many pages are there to manipulate the ranking rather than provide useful information. These can be eliminated or down-weighted via analysis at crawl time or query time. Finally, having so far considered enterprise-scale webs in the Web Track, it is interesting consider ranking with other forms of enterprise information such as mailing list archives and document shares/archives, and a search across a mixture of web and non-web enterprise data.

A All run descriptions

The a description of each run as submitted, sorted as in Figure 2. Each group's best run is marked with a *.

1. **MSRC04C12*** Interleaving submissions MSRC04B1S and MSRC04B2S
2. **MSRC04B2S** Weighted Field BM25 (fields title, body & anchor) optimised on the Named Page 2003 task, with linear adition of non-linear PageRank and URL features. Stemming.
3. **MSRAx2*** relevance propagation + HostRank (more details in Section 2.4 above)
4. **MSRAmixed1** fields weighting + proximity + a new importance named HostRank
5. **MSRAx4** URL match and level + BM25 + HostRank
6. **MSRC04B1S** Weighted Field BM25 (fields title, body & anchor) optimised on the Named Page 2003 task, with linear adition of non-linear PageRank and URL features. No stemming.
7. **MSRAmixed3** BM2500 + Proximity
8. **MSRAx5** relevance propagation + HostRank
9. **MSRC04B1S2** Weighted Field BM25 (fields title, body & anchor) optimised on the Topic Distillation 2003 task, with linear adition of non-linear ClickDistance and URL features. No stemming.
10. **uogWebSelAn*** content and anchor-text retrieval, Porter Stemming, Divergence From Randomness PL2 weighting scheme, URL-length reranking, Selecting between content and anchor-text retrieval, or content with anchor-text and URL-length reranking
11. **UAmsT04MWScb*** CombMNZ (non-normalized, non-weighted) of runs UAmsT04MWinu and UAmsT04MSinu.
12. **uogWebSelAnL** content and anchor-text retrieval, Porter Stemming, Divergence From Randomness PL2 weighting scheme, URL-length reranking, Selecting between content and anchor-text retrieval, or content with anchor-text and URL-length reranking
13. **UAmsT04MSinu** Mixture language model on stemmed full-text, titles, and anchor texts, using both an indegree and URL prior.
14. **UAmsT04MSind** Mixture language model on stemmed full-text, titles, and anchor texts, using an indegree prior.
15. **UAmsT04MWinu** Mixture language model on non-stemmed full-text, titles, and anchor texts, using both an indegree and URL prior.
16. **THUIRmix045*** Word pair weighting based on THUIRmix041.
17. **MSRC04B3S** Weighted Field BM25 (fields title, body & anchor) optimised on the Topic Distillation 2003 task, with linear adition of non-linear ClickDistance No stemming.
18. **THUIRmix044** Query classification with query length and named entity information. TD topics are assigned to THUIRmix042, while the others are retrieved on THUIRmix041.
19. **THUIRmix042** Content retrieval in full text and in-link anchor of Key resource pages. Key resource pages are selected with non-content features using clustering technologies.
20. **ICT04CIIS1AT*** anchor text forward propagation , page title text back propagation, combination of anchor text ,key words ,h1 text etc. ,different pivoted weigth function for different part
21. **ICT04MNZ3** CombMNZ for combination of anchor text retrieval result ,structure info retrieval result and content retrieval result. anchor text forward propagation , page title text back propagation.
22. **uogWebCA** content and anchor text retrieval, Porter Stemming, Divergence From Randomness PL2 weighting scheme
23. **THUIRmix041** Content retrieval in full text and in-link anchor, with a larger weight in fields of Title, head, Bold and first line of page content.
24. **ICT04RULE** rerank the result by some heuristic strategies make use of the url depth,url works,anchkor text, site compression like trick.
25. **uogWebSelL** content and anchor-text retrieval, Porter Stemming, Divergence From Randomness PL2 weighting scheme, URL-length reranking, Selecting between content and anchor-text retrieval, or content with anchor-text and URL-length reranking
26. **THUIRmix043** THUIRmix041 + primary space model weighting in in-link anchor text and contents of Title, head, Bold and first line of page content.
27. **bumW04rdpl*** same as humW04dpl except extra weight for root urls
28. **ICT04CIILC** comparable run with ICT04basic, using a different weighted function for Content text, others just the same as ICT04basic
29. **uogWebCAU150** content and anchor text retrieval, Porter Stemming, Divergence From Randomness PL2 weighting scheme, URL-length reranking
30. **UAmsT04LnuNG** Lnu.ltc run with word n-gram boosting, using document structure and anchor texts.
31. **ICT04basic** vector space content model, baseline for all the runs, using combination of anchor text and some simplest page structure info. not stems,not feedback and classification of queries
32. **SJTUINCMIX3*** BM25
33. **SJTUINCMIX2** Task classification,BM25
34. **MeijiHILw1*** Vector space model. Using anchor text, url-depth and title text. Outdegree reranking.
35. **MeijiHILw3** Vector space model. Using anchor text, url-depth and title text. Outdegree reranking. Query Classified based on last year's queries. Document vector modification by Relevance-based Superimposition Model(RSModel).
36. **SJTUINCMIX1** task classification,BM25,minimal span weighting reRank
37. **MeijiHILw2** Vector space model. Using anchor text, url-depth and title text. Outdegree reranking. Query Classified based on last year's queries.
38. **SJTUINCMIX5** Task classification,BM25,Site Unit
39. **SJTUINCMIX4** Task classification,BM25,PageRank reRank
40. **csiroatnist*** This is a baseline run obtained by submitting the query titles to the Panoptic (CSIRO software) search service at ir.nist.gov. Note that an error with topic 179 resulted in no documents retrieved. To pass the submission checking script, the 30th result for topic 178 was arbitrarily inserted as the first for 179.
41. **bumW04dpl** same as humW04pl except extra weight for lower url depth
42. **MU04web1*** Vector Space Model + Document-centric impact + pagerank + URL depth
43. **bumW04dp** same as humW04dpl except linguistic expansion from stemming disabled
44. **wdf3oks0arr1*** result merging, okapi, simple stemmer, homepage rank boosting
45. **wdf3oks0brr1** result merging, okapi, combo stemmer, homepage rank boosting
46. **wdf3oks0a** result merging, okapi, simple stemmer
47. **MeijiHILw4** Vector space model. Using anchor text, url-depth and title text. Outdegree reranking. Query Classified based on last year's queries.Query expansion using Conceptual Fuzzy Sets(CFS).
48. **wdf3oks0b** result merging, okapi, combo stemmer
49. **humW04pl** same as humW04l except extra weight on properties such as Title and Metadata
50. **VTOK5*** BASELINE
51. **MeijiHILw5** Vector space model. Using anchor text, url-depth and title text. Outdegree reranking. Query Classified based on last year's queries.Query expansion using Conceptual Fuzzy Sets(CFS). Document vector modification by Relevance-based Superimposition Model(RSModel).
52. **MU04web3** Vector Space Model + Document-centric impacts + Pagerank
53. **mpi04web08*** Automatic phrase detection, Anchor text reranking, PageRank, Stemming
54. **mpi04web01** our baseline plain keyword queries from title PageRank Stemming
55. **mpi04web06** Autmatic query expansion + phrase detection PageRank Stemming
56. **mpi04web02** Autmatic query expansion + phrase detection PageRank Stemming
57. **fdwiedf0*** hammingbird algorithm
58. **mpi04web07** Automatic phrase detection, PageRank, Stemming
59. **MU04web5** Vector space model + document-centric impacts
60. **MU04web2** Vector Space Model + Document-centric impacts + URL depth
61. **MU04web4** Vector space model + document-centric impact + pagerank + URL depth
62. **LamMcm1*** Multicriteria analysis Lovins Stemming Kleinberg authority scores
63. **humW04l** plain content search including linguistic expansion from English inflectional stemming
64. **irtbow*** bag of words but with added weighting for query term order and proximity; Lnu.Ltc weighting.
65. **irttil** title only; Lnu.Ltc weighting
66. **fdwiesl0** improved okpai method
67. **irtpbr2** phrase search (not useful for single-term queries); Lnu.Ltc weighting.
68. **fdwiellq1** anchro-text ranking
69. **fdwiellq0** okpai model
70. **XLDBTumba01***
71. **VT2** Ranking tuning using linear fusion
72. **VTTD1** TD tuning
73. **VT1** best trial
74. **VT3** Ranking tuning using linear fusion

Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.

SP 500-261

TREC 2004 Conference Proceedings

NIST

NIST Technical Publications

Periodical

Journal of Research of the National Institute of Standards and Technology—Reports NIST research and development in metrology and related fields of physical science, engineering, applied mathematics, statistics, biotechnology, and information technology. Papers cover a broad range of subjects, with major emphasis on measurement methodology and thc basic technology underlying standardization. Also included from time to time are survey articles on topics closely related to the Institute's technical and scientific programs. Issued six times a year.

Nonperiodicals

Monographs—Major contributions to the technical literature on various subjects related to the Institute's scientific and technical activities.

Handbooks—Recommended codes of engineering and industrial practice (including safety codes) developed in cooperation with interested industries, professional organizations, and regulatory bodies.

Special Publications—Include proceedings of conferences sponsored by NIST, NIST annual reports, and other special publications appropriate to this grouping such as wall charts, pocket cards, and bibliographies.

National Standard Reference Data Series—Provides quantitative data on the physical and chemical properties of materials, compiled from the world's literature and critically evaluated. Developed under a worldwide program coordinated by NIST under the authority of the National Standard Data Act (Public Law 90-396). NOTE:The Journal of Physical and Chemical Reference Data (JPCRD) is published bimonthly for NIST by the American Institute of Physics (AIP). Subscription orders and renewals are available from AIP, P.O. Box 503284, St. Louis, MO63150-3284.

Building Science Series—Disseminates technical information developed at the Institute on building materials, components, systems, and whole structures. The series presents research results, test methods, and performance criteria related to the structural and environmental functions and the durability and safety characteristics of building elements and systems.

Technical Notes—Studies or reports which are complete in themselves but restrictive in their treatment of a subject. Analogous to monographs but not so comprehensive in scope or definitive in treatment of the subject area. Often serve as a vehicle for final reports of work performed at NIST under the sponsorship of other government agencies.

Voluntary Product Standards—Developed under procedures published by the Department of Commerce in Part 10, Title 15, of the Code of Federal Regulations. The standards establish nationally recognized requirements for products, and provide all concerned interests with a basis for common understanding of the characteristics of the products. NIST administers this program in support of the efforts of private-sector standardizing organizations.

Order the following NIST publications—FIPS and NISTIRs—from the National Technical Information Service, Springfield, VA 22161.

Federal Information Processing Standards Publications (FIPS PUB)—Publications in this series collectively constitute the Federal Information Processing Standards Register. The Register serves as the official source of information in the Federal Government regarding standards issued by NIST pursuant to the Federal Property and Administrative Services Act of 1949 as amended, Public Law 89-306 (79 Stat. 1127), and as implemented by Executive Order 11717 (38 FR 12315, dated May 11, 1973) and Part 6 of Title 15 CFR (Code of Federal Regulations).

NIST Interagency or Internal Reports (NISTIR)—The series includes interim or final reports on work performed by NIST for outside sponsors (both government and nongovernment). In general, initial distribution is handled by the sponsor; public distribution is handled by sales through the National Technical Information Service, Springfield, VA 22161, in hard copy, electronic media, or microfiche form. NISTIR's may also report results of NIST projects of transitory or limited interest, including those that will be published subsequently in more comprehensive form.