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

NIST Special Publication 500-272

Information Technology:

The Fifteenth Text Retrieval Conference

TREC 2006

Ellen M. Voorhees
and
Lori P. Buckland,
Editors

Information Technology Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899

er 2007

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, 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-NIST.

Office of the Director
- Baldrige National Quality Program
- Public and Business Affairs
- Civil Rights and Diversity
- International and Academic Affairs

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

Advanced Technology Program
- Economic Assessment
- Information Technology and Electronics
- Chemistry and Life Sciences

Manufacturing Extension Partnership Program
- Center Operations
- Systems Operation
- Program Development

Electronics and Electrical Engineering Laboratory
- Semiconductor Electronics
- Optoelectronics[1]
- Quantum Electrical Metrology
- Electromagnetics

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

NIST Center for Neutron Research

Nanoscale Science and Technology

Chemical Science and Technology Laboratory
- Biochemical Science
- 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
- Materials and Construction Research
- Building Environment
- Fire Research

Information Technology Laboratory
- Mathematical and Computational Sciences[2]
- Advanced Network Technologies
- Computer Security
- Information Access
- Software Diagnostics and Conformance Testing
- Statistical Engineering

[1]At Boulder, CO 80303
[2]Some elements at Boulder, CO

NIST Special Publication 500-272

Information Technology:

The Fifteenth Text Retrieval Conference

TREC 2006

Ellen M. Voorhees
and
Lori P. Buckland,
Editors

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

October 2007

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

National Institute of Standards and Technology
James M. Turner, Acting 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.

National Institute of Standards and Technology Special Publication 500-272
Natl. Inst. Stand. Technol. Spec. Publ. 500-272, 177 pages (October 2007)

Foreword

This report constitutes the proceedings of the 2006 Text REtrieval Conference, TREC 2006, held in Gaithersburg, Maryland, November 14–17, 2006. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Disruptive Technology Office (DTO). Approximately 175 people attended the conference, including representatives from 17 countries. The conference was the fifteenth in an ongoing 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.

I gratefully acknowledge the tremendous work of the TREC program committee and the track coordinators.

Ellen Voorhees
September 24, 2007

TREC 2006 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
Bill Hersh, Oregon Health & Science University
David Lewis, David Lewis Consulting
John Prager, IBM
Steve Robertson, Microsoft
Mark Sanderson, University of Sheffield
Ian Soboroff, NIST
Karen Sparck Jones, University of Cambridge, UK
Richard Tong, Tarragon Consulting
Ross Wilkinson, CSIRO

TREC 2006 Proceedings

Overview Papers

Other Papers

(Contents of these papers are found on the TREC 2006 Proceedings CD.)

Appendix

(Contents of the Appendix are found on the TREC 2006 Proceedings CD.)

Papers: Alphabetical by Organization

(Contents of these papers are found on the TREC 2006 Proceedings CD.)

Henggeler Computer Consultants
QACTIS Enhancements in TREC QA 2006

Humboldt Universität zu Berlin
Highly Scalable Discriminative Spam Filtering

IBM Haifa Research Lab
Juru at TREC 2006: TAAT versus DAAT in the Terabyte Track

IBM T.J. Watson Research Center
IBM in TREC 2006 Enterprise Track

Illinois Institute of Technology
IIT TREC 2006: Genomics Track

Indiana University, Bloomington
WIDIT in TREC 2006 Blog Track

Institute for Infocomm Research
I2R at TREC 2006 Genomics Track

ITC-irst
Reconstructing DIOGENE: ITC-irst at TREC 2006

Johns Hopkins Applied Physics Laboratory
QACTIS Enhancements in TREC QA 2006

Jozef Stefan Institute (and Klika)
Towards Practical PPM Spam Filtering: Experiments for the TREC 2006 Spam Track

Kobe University
Opinion Retrieval Experiments Using Generative Models: Experiments for the TREC 2006 Blog Track

Kyoto University
Combining Vector-Space and Word-Based Aspect Models for Passage Retrieval

L3S Research Center
L3S Research Center at TREC 2006 Enterprise Track

Language Computer Corporation
Question Answering with LCC's CHAUCER at TREC 2006

A Temporally Enhanced PowerAnswer in TREC 2006

LexiClone
LexiClone Lexical Cloning Systems

Macquarie University
AnswerFinder at TREC 2006

Max-Planck-Institut für Informatik
IO-Top-k at TREC 2006: Terabyte Track

Microsoft Research
Window-based Enterprise Expert Search

Microsoft Research Cambridge
Overview of the TREC 2006 Enterprise Track

Microsoft Research Asia
Research on Expert Search at Enterprise Track of TREC 2006

MIT Computer Science and Artificial Intelligence Laboratory
Question Answering Experiments and Resources

The MITRE Corporation
MITRE's Qanda at TREC 15

National Archives and Records Administration
TREC 2006 Legal Track Overview

National Institute of Informatics
Opinion Retrieval Experiments Using Generative Models: Experiments for the TREC 2006 Blog Track

National Institute of Standards and Technology (NIST)
Overview of TREC 2006

Overview of the TREC 2006 Blog Track

Overview of the TREC 2006 Enterprise Track

Overview of the TREC 2006 Question Answering Track

National Library of Medicine
Finding Relevant Passages in Scientific Articles: Fusion of Automatic Approaches vs. an Interactive Team Effort

National Taiwan University
NTU at TREC 2006 Genomics Track

NEC Laboratories America
The Splog Detection Task and A Solution Based on Temporal and Link Properties

Northeastern University
The Hedge Algorithm for Metasearch at TREC 2006

Open Text Corporation
Experiments with the Negotiated Boolean Queries of the TREC 2006 Legal Discovery Track

The Open University
The Open University at TREC 2006 Enterprise Track Expert Search Task

Oregon Health & Science University
TREC 2006 Genomics Track Overview

Combining Lexicon Expansion, Information Retrieval, and Cluster-Based Ranking for Biomedical Question Answering

Peking University
Tianwang at TREC 2006 QA Track

Peking University at the TREC 2006 Terabyte Track

Purdue University
Knowledge Transfer and Opinion Detection in the TREC 2006 Blog Track

Combining Multiple Resources, Evidences and Criteria for Genomic Information Retrieval

Queen Mary University of London
Solving the Enterprise TREC Task with Probabilistic Data Models

Ricoh Software Research Center
Ricoh Research at TREC 2006: Enterprise Track

RMIT University
RMIT University at TREC 2006: Terabyte Track

The Robert Gordon University
The Robert Gordon University

Saarland University
The Alyssa System at TREC 2006: A Statistically Inspired Question Answering System

Shanghai Jiao Tong University
Research on Expert Search at Enterprise Track of TREC 2006

State University of New York at Buffalo
UB at TREC Genomics 2006: Using Passage Retrieval and Pre-Retrieval Query Expansion for Genomics IR

Finding Relevant Passages in Scientific Articles: Fusion of Automatic Approaches vs. an Interactive Team Effort

St. Petersburg State University
SOPHIA in Enterprise Track

Swiss Institute of Bioinformatics
Report on the TREC 2006 Experiment: Genomics Track

TNO
Concept-Based Document Retrieval for Genomics Literature

Tokyo Institute of Technology
TREC 2006 Question Answering Experiments at Tokyo Institute of Technology

TrulyIntelligent Technologies
TREC 2006 Q&A Factoid: TI Experience

Tufts University
Spam Filtering Using Inexact String Matching in Explicit Feature Space with Online Linear Classifiers

University of Albany SUNY
ILQUA at TREC 2006

University of Amsterdam
Overview of the TREC 2006 Blog Track

Multiple Ranking Strategies for Opinion Retrieval in Blogs: The University of Amsterdam at the 2006 TREC Blog Track

Language Models for Enterprise Search: Query Expansion and Combination of Evidence

Experiments with Document and Query Representations for a Terabyte of Text

Expanding Queries Using Multiple Resources

University of Arkansas at Little Rock
UALR at TREC: Blog Track

University of California, Berkeley
BioText Team Report for the TREC 2006 Genomics Track

University of California Santa Cruz
UCSC on REC 2006 Blog Opinion Mining

University of Colorado Health Sciences Center
Concept Recognition, Information Retrieval, and Machine Learning in Genomics Question Answering

Università degli Studi di Milano
MG4J at TREC 2006

University of Edinburgh
Experiments at the University of Edinburgh for the TREC 2006 QA Track

University of Geneva
Report on the TREC 2006 Experiment: Genomics Track

University of Glasgow
Overview of the TREC 2006 Blog Track

University of Glasgow at TREC 2006: Experiments in Terabyte and Enterprise Tracks with Terrier

University of Guelph
Passage Retrieval by Shrinkage of Language Models

University Hospital of Geneva
Finding Relevant Passages in Scientific Articles: Fusion of Automatic Approaches vs. an Interactive Team Effort

Report on the TREC 2006 Experiment: Genomics Track

University of Illinois Chicago
UIC at TREC 2006 Blog Track

A Concept-Based Framework for Passage Retrieval at Genomics

University of Illinois at Urbana-Champaign
Language Models for Expert Finding--UIUC TREC 2006 Enterprise Track Experiments

Robust Pseudo Feedback Estimation and HMM Passage Extraction: UIUC at TREC 2006 Genomics Track

The University of Iowa
Extraction of Document Structure for Genomics Documents

Universität Karlsruhe
The Ephyra QA System at TREC 2006

University of Limerick
Question Answering Using the DLT System at TREC 2006

University of Ljubljana
Towards Practical PPM Spam Filtering: Experiments for the TREC 2006 Spam Track

University of Maryland, College Park
TREC 2006 Legal Track Overview

Overview of the TREC 2006 Question Answering Track

Finding Relevant Passages in Scientific Articles: Fusion of Automatic Approaches vs. an Interactive Team Effort

TREC 2006 at Maryland: Blog, Enterprise, Legal and QA Tracks

University of Maryland, Baltimore County
The BlogVox Opinion Retrieval System

University of Waterloo
TREC 2006 Spam Track Overview

The TREC 2006 Terabyte Track

Index Pruning and Result Reranking: Effects on Ad Hoc Retrieval and Named-Page Finding

In Enterprise Search: Methods to Identify Argumentative Discussions and to Find Topical Experts

Identifying Relationships Between Entities in Text for Complex Interactive Question Answering Task

University of Wisconsin, Madison
Ranking Biomedical Passages for Relevance and Diversity: University of Wisconsin, Madison at TREC Genomics 2006

U.S. Department of Defense
QACTIS Enhancements in TREC QA 2006

Wuhan University
Window-based Enterprise Expert Search

Weill Medical College of Cornell University
Twease at TREC 2006: Breaking and Fixing BM25 Scoring With Query Expansion, A Biologically Inspired Double Mutant Recovery

York University
York University at TREC 2006: Enterprise Email Discussion Search

York University at TREC 2006: Genomics Track

York University at TREC 2006: Legal Track

Papers: Organized by Track

(Contents of these papers are found on the TREC 2006 Proceedings CD.)

Blog

Carnegie Mellon
Knowledge Transfer and Opinion Detection in the TREC 2006 Blog

Chinese Academy of Sciences
Combining Language Model with Sentiment Analysis for Opinion Retrieval of Blog-Post

CSIRO ICT Team
Using Semantic Relations with World Knowledge for Question Answering

Indiana University, Bloomington
WIDIT in TREC 2006 Blog Track

Kobe University
Opinion Retrieval Experiments Using Generative Models: Experiments for the TREC 2006 Blog Track

National Institute of Informatics
Opinion Retrieval Experiments Using Generative Models: Experiments for the TREC 2006 Blog Track

National Institute of Standards and Technology (NIST)
Overview of the TREC 2006 Blog Track

NEC Laboratories America
The Splog Detection Task and A Solution Based on Temporal and Link Properties

Purdue University
Knowledge Transfer and Opinion Detection in the TREC 2006 Blog

The Robert Gordon University
The Robert Gordon University

University of Amsterdam
Overview of the TREC 2006 Blog Track

Multiple Ranking Strategies for Opinion Retrieval in Blogs: The University of Amsterdam at the 2006 TREC Blog Track

University of Arkansas at Little Rock
UALR at TREC: Blog Track

University of California Santa Cruz
UCSC on REC 2006 Blog Opinion Mining

University of Glasgow
Overview of the TREC 2006 Blog Track

University of Illinois Chicago
UIC at TREC 2006 Blog Track

University of Maryland, College Park
TREC 2006 at Maryland: Blog, Enterprise, Legal and QA Tracks

University of Maryland, Baltimore County
The BlogVox Opinion Retrieval System

Univerity of North Carolina
Opinion Retrieval Experiments Using Generative Models: Experiments for the TREC 2006 Blog Track

Università di Pisa
Blog Mining Through Opinionated Words

Genomics

Arizona State University
ASU at TREC 2006 Genomics Track

Carnegie Mellon University
Combining Multiple Resources, Evidences and Criteria for Genomic Information Retrieval

Concordia University
BioKI, A General Literature Navigation System at TREC Genomics 2006

Dalian University of Technology
DUTIR at TREC 2006: Genomics and Enterprise Tracks

Free University Amsterdam
Expanding Queries Using Multiple Resources

Fudan University
Using Profile Matching and Text Categorization for Answer Extraction in TREC Genomics

Illinois Institute of Technology
IIT TREC 2006: Genomics Track

Institute for Infocomm Research
I2R at TREC 2006 Genomics Track

Kyoto University
Combining Vector-Space and Word-Based Aspect Models for Passage Retrieval

National Library of Medicine
Finding Relevant Passages in Scientific Articles: Fusion of Automatic Approaches vs. an Interactive Team Effort

Oregon Health & Sciences University
Combining Lexicon Expansion, Information Retrieval, and Cluster-Based Ranking for Biomedical Question Answering

Purdue University
Combining Multiple Resources, Evidences and Criteria for Genomic Information Retrieval

State University of New York at Buffalo
Finding Relevant Passages in Scientific Articles: Fusion of Automatic Approaches vs. an Interactive Team Effort

UB at TREC Genomics 2006: Using Passage Retrieval and Pre-Retrieval Query Expansion for Genomics IR

Swiss Institute of Bioinformatics
Report on the TREC 2006 Experiment: Genomics Track

University Amsterdam
Expanding Queries Using Multiple Resources

University of California, Berkeley
BioText Team Report for the TREC 2006 Genomics Track

University of Colorado Health Sciences Center
Concept Recognition, Information Retrieval, and Machine Learning in Genomics Question Answering

University of Geneva
Report on the TREC 2006 Experiment: Genomics Track

University of Guelph
Passage Retrieval by Shrinkage of Language Models

University Hospital of Geneva
Finding Relevant Passages in Scientific Articles: Fusion of Automatic Approaches vs. an Interactive Team Effort

Report on the TREC 2006 Experiment: Genomics Track

University of Illinois at Chicago
A Concept-Based Framework for Passage Retrieval at Genomics

University of Illinois at Urbana-Champaign
Robust Pseudo Feedback Estimation and HMM Passage Extraction: UIUC at TREC 2006 Genomics Track

University of Maryland, College Park
Finding Relevant Passages in Scientific Articles: Fusion of Automatic Approaches vs. an Interactive Team Effort

University of Massachusetts, Amherst
UMass Genomics 2006: Query-Biased Pseudo Relevance Feedback

The University of Melbourne
Combining Vector-Space and Word-Based Aspect Models for Passage Retrieval

University of Neuchatel
Report on the TREC 2006 Genomics Experiment

University of Wisconsin, Madison
Ranking Biomedical Passages for Relevance and Diversity: University of Wisconsin, Madison at TREC Genomics 2006

Weill Medical College of Cornell University
Twease at TREC 2006: Breaking and Fixing BM25 Scoring With Query Expansion, A Biologically Inspired Double Mutant Recovery Experiment

York University
York University at TREC 2006: Genomics Track

Enterprise

Beijing University of Posts and Telecommunications
BUPT at TREC 2006: Enterprise Track

Case Western Reserve University
Case Western Reserve University at the TREC 2006 Enterprise Track

Chinese Academy of Sciences
Social Network Structure Behind the Mailing Lists: ICT-IIIS at TREC 2006 Expert Finding Track

City University
Window-based Enterprise Expert Search

City University London
Window-based Enterprise Expert Search

CWI
MonetDB/X100 at the 2006 TREC Terabyte Track

Overview of the TREC 2006 Enterprise Track

Dalian University of Technology
DUTIR at TREC 2006: Genomics and Enterprise Tracks

Fudan University Shanghai
Judging Expertise--WIM at Enterprise

IBM T.J. Watson Research Center
IBM in TREC 2006 Enterprise Track

L3S Research Center
L3S Research Center at TREC 2006 Enterprise Track

Microsoft Cambridge
Overview of the TREC 2006 Enterprise Track

Microsoft Research
Window-based Enterprise Expert Search

Microsoft Research Asia
Research on Expert Search at Enterprise Track of TREC 2006

National Institute of Standards and Technology
Overview of the TREC 2006 Enterprise Track

The Open University
The Open University at TREC 2006 Enterprise Track Expert Search Task

Queen Mary University of London
Solving the Enterprise TREC Task with Probabilistic Data Models

Ricoh Software Research Center
Ricoh Research at TREC 2006: Enterprise Track

Shanghai Jiao Tong University
Research on Expert Search at Enterprise Track of TREC 2006

St. Petersburg State University
SOPHIA in Enterprise Track

University of Amsterdam
Language Models for Enterprise Search: Query Expansion and Combination of Evidence

University of Glasgow
University of Glasgow at TREC 2006: Experiments in Terabyte and Enterprise Tracks with Terrier

University of Illinois at Urbana-Champaign
Language Models for Expert Finding--UIUC TREC 2006 Enterprise Track Experiments

University of Maryland, College Park
TREC 2006 at Maryland: Blog, Enterprise, Legal and QA Tracks

University of Massachusetts, Amherst
UMass at TREC 2006: Enterprise Track

University of Pittsburgh
IBM in TREC 2006 Enterprise Track

Pitt at TREC 2006: Identifying Experts via Email Discussions

University of Ulster
SOPHIA in Enterprise Track

University of Waterloo
In Enterprise Search: Methods to Identify Argumentative Discussions and to Find Topical Experts

Wuhan University, China
Window-based Enterprise Expert Search

York University
York University at TREC 2006: Enterprise Email Discussion Search

Legal

David D. Lewis Consulting
TREC 2006 Legal Track Overview

National Archives and Records Administration
TREC 2006 Legal Track Overview

Open Text Corporation
Experiments with the Negotiated Boolean Queries of the TREC 2006 Legal Discovery Track

University of Maryland, College Park
TREC 2006 Legal Track Overview

TREC 2006 at Maryland: Blog, Enterprise, Legal and QA Tracks

University of Missouri, Kansas City
Experiments with Query Expansion at TREC 2006 Legal Track

York University
York University at TREC 2006: Legal Track

Question Answering

Bilkent University
Identifying Relationships Between Entities in Text for Complex Interactive Question Answering Task

Concordia University
Concordia University at the TREC 15 QA Track

Dalhousie University
DalTREC 2006 QA System Jellyfish: Regular Expressions Mark-and-Match Approach to Question Answering

Diggery
Question Answering by Diggery at TREC 2006

Fudan University
FDUQA on TREC 2006 QA Track

Harbin Institute of Technology
InsunQA06 on QA Track of TREC 2006

ITC-irst
Reconstructing DIOGENE: ITC-irst at TREC 2006

Language Computer Corporation
Question Answering with LCC's CHAUCER at TREC 2006

A Temporally Enhanced PowerAnswer in TREC 2006

LexiClone
LexiClone Lexical Cloning Systems

Macquarie University
AnswerFinder at TREC 2006

MIT Computer Science and Artificial Intelligence Laboratory
Question Answering Experiments and Resources

The MITRE Corporation
MITRE's Qanda at TREC 15

National Institute of Standards and Technology
Overview of the TREC 2006 Question Answering Track

Peking University
Tianwang at TREC 2006 QA Track

Saarland University
The Alyssa System at TREC 2006: A Statistically Inspired Question Answering System

Tokyo Institute of Technology
TREC 2006 Question Answering Experiments at Tokyo Institute of Technology

TrulyIntelligent Technologies
TREC 2006 Q&A Factoid: TI Experience

University of Albany SUNY
ILQUA at TREC 2006

University of Edinburgh
Experiments at the University of Edinburgh for the TREC 2006 QA Track

Universität Karlsruhe
The Ephyra QA System at TREC 2006

University of Limerick
Question Answering Using the DLT System at TREC 2006

University of Maryland, College Park
TREC 2006 at Maryland: Blog, Enterprise, Legal and QA Tracks

Overview of the TREC 2006 Question Answering Track

University of Massachusetts, Amherst
UMass at TREC ciQA 2006

University of North Carolina, Chapel Hill
Overview of the TREC 2006 Question Answering Track

University of Rome "La Sapienza"
The "La Sapienza" Question Answering System at TREC 2006

University of Sheffield
The University of Sheffield's TREC 2006 Q&A Experiments

University of Strathclyde
Contextual Information and Assessor Characteristics in Complex Question Answering

The University of Washington
The University of Washington's UWclmaQA System

University of Waterloo
Identifying Relationships Between Entities in Text for Complex Interactive Question Answering Task

Spam

Beijing University of Posts and Telecommunications
BUPT at TREC 2006: Spam Track

CRM114 Team
Seven Hypothesis about Spam Filtering

Fidelis Assis
OSBF-Lua - A Text Classification Module for Lua: The Importance of the Training Method

Harbin Institute of Technology
SVM-Based Spam Filter with Active and Online Learning

Humboldt Universität zu Berlin
Highly Scalable Discriminative Spam Filtering

Jozef Stefan Institute
Towards Practical PPM Spam Filtering: Experiments for the TREC 2006 Spam Track

Tufts University
Spam Filtering Using Inexact String Matching in Explicit Feature Space with Online Linear Classifiers

University of Ljubljana
Towards Practical PPM Spam Filtering: Experiments for the TREC 2006 Spam Track

University of Waterloo
TREC 2006 Spam Track Overview

Terabyte

Arctic Region Supercomputing Center
Partitioning the Gov2 Corpus by Internet Domain Name: A Result-Set Merging Experiment

Chinese Academy of Sciences
PSM: A New Re-Ranking Algorithm for Named Page

CWI
MonetDB/X100 at the 2006 TREC Terabyte Track

Dublin City University
Dublin City University at the TREC 2006 Terabyte Track

Ecole Nationale Superieure des Mines de Saint Etienne
Fuzzy Term Proximity With Boolean Queries at 2006 TREC Terabyte Task

IBM Haifa Research Lab
Juru at TREC 2006: TAAT versus DAAT in the Terabyte Track

Max-Planck-Institut für Informatik
IO-Top-k at TREC 2006: Terabyte Track

NIST
The TREC 2006 Terabyte Track

Northeastern University
The Hedge Algorithm for Metasearch at TREC 2006

Peking University
Peking University at the TREC 2006 Terabyte Track

RMIT University
RMIT University at TREC 2006: Terabyte Track

Abstract

This report constitutes the proceedings of the 2006 Text REtrieval Conference, TREC 2006, held in Gaithersburg, Maryland, November 14–17, 2006. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Disruptive Technology Office (DTO). TREC 2006 had 107 participating groups including participants from 17 countries.

TREC 2006 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: search in support of legal discovery of electronic documents, search within and between blog postings, question answering, detecting spam in an email stream, enterprise search, search on (almost) terabyte-scale document sets, and search in the genomics domain.

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 2006 are contained on the disk included in the volume. The TREC 2006 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 2006

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

1 Introduction

The fifteenth Text REtrieval Conference, TREC 2006, was held at the National Institute of Standards and Technology (NIST) 14 to 17 November 2006. The conference was co-sponsored by NIST and the Disruptive Technology Office (DTO). TREC 2006 had 107 participating groups from 17 different countries. Table 2 at the end of the paper lists the participating groups.

TREC 2006 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 2006 contained seven areas of focus called "tracks". Five of the tracks ran in previous TRECs and explored tasks in question answering, detecting spam in an email stream, enterprise search, search on (almost) terabyte-scale document sets, and information access within the genomics domain. The two new tracks explored blog search and providing support for legal discovery of electronic documents.

There were two main themes in TREC 2006 that were supported by these different tracks. The first theme was exploring broader information contexts than in previous TRECs. This was accomplished by exploring both different document genres and different retrieval tasks. Traditional TREC document genres of newswire (in the QA track) and web pages (in the terabyte track) were still used, but these were joined by blogs (blog track), email (enterprise and spam tracks), corporate repositories (enterprise and legal tracks), and scientific documents (genomic and legal tracks). Retrieval tasks examined included ad hoc search (terabyte, enterprise-discussion, legal, genomics), known-item search (terabyte), classification (spam), specific response (QA, genomics, enterprise-expert), and opinion finding (blog). The second theme of the conference was a focus on creating new evaluation methodologies. These efforts included examining how to make fair comparisons when using massive data sets (terabyte and legal tracks), assessing the quality of a specific response (genomics, QA), balancing realism and privacy protection in experimental design (spam, enterprise), and constructing protocols for efficiency benchmarking in a distributed setting (terabyte).

This paper serves as an introduction to the research described in detail in the remainder of the proceedings. 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 blog post, an email message, or an invoice.

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. The main task in the terabyte track and the legal track task are examples of ad hoc search 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 task in the terabyte track is an example of a known-item search task.

In a *categorization* task, the system is responsible for assigning a document to one or more categories from among a given set of categories. Deciding whether a given mail message is spam is one example of a categorization task, while the opinion search task in the blog track and the discussion search task in the enterprise track are mixtures of ad hoc and categorization tasks.

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. In addition, the passage retrieval focus in the genomics track is a move toward question answering, and the expert-finding task in the enterprise track is a kind of question answering task in that the system response to an expert-finding search is a set of people, not documents.

2.1 Test collections

Text retrieval has a long history of using retrieval experiments on test collections to advance the state of the art [4, 8], and TREC continues this tradition. A test collection is an abstraction of an operational retrieval

```
<num> Number:  758
<title> Sugar tariff-rate quotas
<desc> Description:  Describe the nature and history of sugar
tariff-rate quotas in the United States.
<narr> Narrative:  Documents describing the system, its history and how
it works are relevant.  Proposed changes to the system or new agreements
explaining how it works are relevant.  Listings of current allocations
are not relevant.
```

Figure 1: A sample TREC 2006 topic from the terabyte track test set.

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. We call the result of a retrieval system executing a task on a test collection a run.

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 initial TREC test collections contain 2 to 3 gigabytes of text and 500,000 to 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 in TREC 2004 to investigate both retrieval and evaluation issues associated with collections significantly larger than 2 gigabytes of text.

While the initial TREC document sets consisted mostly of newspaper or newswire articles, later document sets have included recordings of speech, web pages, scientific documents, blog posts, email messages, and so forth. In each case, high-level structures within each document are tagged using SGML or XML, and each document is assigned an unique identifier called the DOCNO. In keeping of the spirit of realism, text is kept as close to the original as possible. No attempt is 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—though some tracks don't use topics at all (e.g., spam) or use different formats to support the track (e.g., legal). An example topic taken from this year's terabyte 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 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 ad hoc 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 [6]. 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 [9].

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 [7] 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. Pooling is valid when enough relevant documents are found to make the resulting judgment set approximately complete and unbiased.

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 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 $X \times$ *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 [12]. 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 [10]. 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].

The leave-out-uniques (LOU) test can fail to indicate a problem with a collection if all the runs that contribute to the pool share a common bias—preventing such a common bias is why a diverse run set is needed for pool construction. While it is not possible to prove that no common bias exists for a collection, no common bias has been demonstrated for any of the TREC collections until recently. When pools are shallow *relative to the number of documents in the collection*, the sheer number of documents of a certain type fill up the pools to the exclusion of other types of documents. In particular, otherwise diverse retrieval methodologies will all rank documents that have lots of topic title words before documents containing fewer topic title words since topic title words are specifically chosen to be good content indicators. To produce an unbiased, reusable collection, traditional pooling requires sufficient room in the pools to exhaust the spate of title-word documents and allow documents that are not title-word-heavy to enter the pool [2]. But large document sets such as the one used in the terabyte track include so many documents containing topic title words that traditional pooling requires pools that are much far too large to be affordable to judge. One of the goals for the terabyte track is to investigate new pooling strategies to build reusable, fair collections at a reasonable cost despite collection size.

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 (number-retrieved-and-relevant/number-retrieved), while recall is the proportion of relevant documents that are retrieved (number-retrieved-and-relevant/number-relevant). 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 at ten documents retrieved less than 1.0 regardless of how the documents are ranked. Similarly, a topic with more than ten relevant documents must have a recall score at ten documents retrieved less than 1.0. 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 interpolated recall-precision curve and mean average precision (non-interpolated) 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 (MAP) 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, existing evaluation measures have been adapted and new evaluation measures have been devised. The details of the evaluation methodology used in a particular track are described in the track's overview paper.

3 TREC 2006 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

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

	TREC														
Track	'92	'93	'94	'95	'96	'97	'98	'99	'00	'01	'02	'03	'04	'05	'06
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										
Merging				3	3										
Filtering				4	7	10	12	14	15	19	21				
Chinese					9	12									
NLP					4	2									
Speech						13	10	10	3						
XLingual						13	9	13	16	10	9				
High Prec						5	4								
VLC							7	6							
Query							2	5	6						
QA								20	28	36	34	33	28	33	31
Web								17	23	30	23	27	18		
Video										12	19				
Novelty											13	14	14		
Genomics												29	33	41	30
HARD												14	16	16	
Robust												16	14	17	
Terabyte													17	19	21
Enterprise														23	25
Spam														13	9
Legal															6
Blog															16
Participants	22	31	33	36	38	51	56	66	69	87	93	93	103	117	107

of participants, but has also created a smaller common base of experience among participants since each participant tends to submit runs to a smaller percentage of the tracks.

This section describes the tasks performed in the TREC 2006 tracks. See the track reports later in these proceedings for a more complete description of each track.

3.1 The blog track

The blog track is a new track in TREC 2006. Its purpose is to explore information seeking behavior in the blogosphere, in particular to discover the similarities and differences between blog search and other types of search. The track contained two tasks, an open task and an opinion retrieval task. Participants in the open task defined their own retrieval task and evaluation strategy using the blog corpus. These were pilot evaluations to inform the discussion of the track's future. The opinion retrieval task was a common task with topic development and relevance judgments performed at NIST.

The blog corpus was collected over a period of 11 weeks from December 2005 through February 2006. It consists of a set of uniquely-identified XML feeds and the corresponding blog posts in HTML. A "document" in the collection (for the purposes of the opinion task) is a single blog post plus all of its associated comments as identified by a Permalink. The collection is a large sample of the blogosphere as it existed in early 2006 that retains all of the gathered material including spam, potentially offensive content, and some non-blogs such as RSS feeds.

In the opinion task, systems were to locate blog posts that expressed an opinion about a given target. Targets included people, organizations, locations, product brands, technology types, events, literary works, etc. For example, three of the test set topics asked for opinions regarding the Macbook Pro, Jon Stewart, and super bowl ads. Targets were drawn from a log of queries submitted to BlogPulse. The query from the log was used as the title field of the topic statement and the NIST assessor created the remaining parts of the topic statement.

While the systems' task was to retrieve posts expressing an opinion of the target without regard to the polarity of the opinion, the relevance assessments made for the track did differentiate among different types of posts to provide useful training data for future tasks. A post could remain unjudged if it was clear from the URL or header that the post contains offensive content. If the content was judged, it was marked with exactly one of: irrelevant (not on-topic), relevant but not opinionated (on-topic but no opinion expressed), relevant with negative opinion, relevant with mixed opinion, or relevant with positive opinion.

Fourteen groups participated in the blog opinion task, and an additional two groups participated in the open task. The primary measure used in the track was MAP when treating a document as relevant if it was both on-topic and opinionated. Runs were also evaluated using just on-topic as the definition of relevant. The correlation between the system rankings produced by the two definitions of relevant was high (a Spearman's ρ of 0.97 and a Kendall's τ of .88), suggesting that whether or not a document was on-topic dominated the retrieval results. A baseline run (created after relevance judging was complete) produced by the University of Glasgow's Terrier system with no opinion-specific processing was more effective than any of the submitted systems using either of the definitions of relevant. Thus more work is required to be able to separate opinionated posts from on-topic posts.

3.2 The enterprise track

The enterprise track started in TREC 2005. The purpose of the track is to study enterprise search: satisfying a user who is searching the data of an organization to complete some task. Enterprise data generally consists of diverse types such as published reports, intranet web sites, and email, and the goal is to have search systems deal seamlessly with the different data types.

The document set used in both years of the track was the W3C Test collection (see `http://research.microsoft.com/users/nickcr/w3c-summary.html`). This collection, created by Nick Craswell, was created from a crawl of the World-Wide Web Consortium web site and includes email discussion lists, web pages, and the extracted text from documents in various formats (such as pdf, postscript, Word, Powerpoint, etc.).

The track contained two tasks, a discussion search task and a search-for-experts task. A total of twenty-five groups participated in the enterprise track.

In the discussion search task the systems were to retrieve the set of messages in the email lists that provided pro/con arguments for a particular choice such as "html vs. xhtml". The task was specifically focused on finding arguments for or against a decision rather than simply finding information about the topic. The motivation for the task is to assist users in understanding why a particular decision has been made.

The runs were evaluated both when relevance was defined simply as being on-topic as well as when relevance was defined as containing a pro/con argument. With a few exceptions including a manual run, the relative effectiveness of the runs was largely the same in both cases. Indeed, a more detailed look at the document rankings (see the track overview paper for details) showed that most runs did not consistently retrieve documents containing an argument earlier than documents that were simply on-topic. Thus, more

work is needed to develop argument detectors.

The motivation for the expert-finding task is being able to determine who to contact regarding a particular matter in a large organization. As operationalized in the track task, the expert search mines an organization's documents to create profiles of its people. Systems returned a ranked list of person-ids and a set of supporting documents per person in response to a topic such as "ontology engineering". Systems were given a mapping between names and person-ids of W3C members. The supporting documents were a set of up to 20 documents that the system believed demonstrated why the person was an expert on the topic. Topic creation and relevance assessments were performed by the track participants.

The better expert-finding runs had a mean reciprocal rank (MRR) score greater than 0.9 showing that those systems were generally able to return a true expert at rank one. Corresponding P(10) scores were approximately 0.7 showing that the majority of candidate experts suggested by those runs were in fact experts.

3.3 The genomics track

The goal of genomics track is to provide a forum for evaluation of information access systems in the genomics domain. It is the first TREC track devoted to retrieval within a specific domain, and thus a subgoal of the track is to explore how exploiting domain-specific information improves access. The TREC 2006 track consisted of a single passage retrieval task, though that task was evaluated in a number of different ways to explore a variety of facets. The task was motivated by the observation that the best response for a biomedical literature search is frequently a direct answer to the question, but with the answer placed in context and linking to original sources.

The document set used in the track was a set of full-text articles from several biomedical journals which were made available to the track by Highwire Press. The documents retain the full formatting information (in HTML) and include tables, figure captions, and the like. The test set contains 162,259 documents from 49 journals and is about 12.3 GB of HTML. A passage is defined to be any contiguous span of text that does not include an HTML paragraph token (<p> or <\p>). Systems returned a ranked list of passages in response to a topic where passages were specified by byte offsets from the beginning of the document.

The topics were derived from the topics used in the TREC 2005 track. The form of the topic was a natural language question, though these were created using a set of "generic topic templates" such as *Find articles describing the role of a gene involved in a given disease*. The test set contained 28 questions, seven questions each from four templates.

Relevance judgments were made by 10 people with expertise in the domain. The judgment process involved several steps to enable system responses to be evaluated at different levels of granularity. Passages from different runs were pooled, using the maximum extent of a passage as the unit for pooling. (The maximum extent of a passage is the contiguous span between paragraph tags that contains that passage, assuming a virtual paragraph tag at the beginning and end of each document.) Judges decided whether a maximum span was relevant (contained an answer to the question), and, if so, marked the actual extent of the answer in the maximum span. In addition, the assessor assigned one or more MeSH terms to that passage as the definition of the *aspect* that the passage pertained to. A maximum span could contain multiple answer passages; the same aspect could be covered by multiple answer passages and a single answer passage could pertain to multiple aspects.

Using these relevance judgments, runs were then evaluated at the document, passage, and aspect levels. A document is considered relevant if it contains a relevant passage, and it is considered retrieved if any of its passages are retrieved. The document level evaluation was a traditional ad hoc retrieval task (when

all subsequent retrievals of a document after the first were ignored). Passage- and aspect-level evaluation was based on the corresponding judgments. Aspect-level evaluation is a measure of the diversity of the retrieved set in that it rewards systems that are able to find more different aspects. Passage-level evaluation is a measure of how well systems are able to find the particular information within a document that answers the question.

The genomics track had 30 participants. The passage-level task is apparently a difficult task as evaluation scores for this task were generally low. Effectiveness for both the aspect and document levels was much better, suggesting that the difficulty for the passage level is in finding the appropriate extent of the required information.

3.4 The legal track

The legal track was a new track in 2006. It focused on a specific aspect of retrieval in the legal domain, that of meeting the needs of lawyers to engage in effective discovery of digital documents. Currently, it is common for the two sides involved in litigation to negotiate a Boolean expression that defines the set of documents that are then examined by humans to determine which are responsive to a discovery request. The goal of the track is to evaluate the effectiveness of other search technologies in facilitating this process.

From the retrieval perspective, the task in the track was an ad hoc search task using a set of hypothetical complaints and requests for the production of documents as topics. The document set used in the track was the IIT Complex Document Information Processing collection, which consists of approximately seven million documents drawn from the Legacy Tobacco Document Library hosted by the University of California at San Francisco. These documents were made public during various legal cases involving US tobacco companies and contain a wide variety of document genres typical of large enterprise environments. A document in the collection consists of the optical character recognition (OCR) output of a scanned original plus a metadata record.

The production requests used as topics were developed for the track by lawyers and were designed to simulate the kinds of requests used in current practice. Each production request includes a broad complaint that lays out the background for several requests and one specific request for production of documents. The topic statement also includes a negotiated Boolean query for each specific request. Systems could use the negotiated Boolean query in any way they saw fit (including ignoring it completely) for the TREC runs. Stephen Tomlinson of Open Text (Hummingbird) ran the track's reference run, which consisted of running just the negotiated Boolean query for each topic.

The relevance assessments were made by legal professionals who followed their typical work practices. Pools were created using traditional pooling for the TREC submissions received from the six participating groups plus a stratified sample of the baseline Boolean run. In addition, the track organizers arranged for a professional searcher familiar with the document collection to (manually) produce a set of approximately 100 documents for each topic that the searcher expected to be relevant to the topic and unlikely to be retrieved by the other methods. These documents were also added to the pools.

To understand how ranked retrieval approaches can assist discovery, it is necessary to compare ranked retrieval results to the results obtained by the negotiated Boolean queries. Thus, one of the goals of the track was the development of an evaluation methodology that provides for the fair comparison of such runs on a large document set where only a sample of documents is judged. This is a very complicated issue that this first running of the track has just begun to address. In the interim, one measure used in the track was R-precision, a measure that probably favors ranked retrieval runs since the "first" R documents is not well-defined in a pure Boolean run. However, each of the Boolean runs submitted to the track including

the reference run were ranked in some fashion after the Boolean constraint was applied, so R-precision is defined for the track runs. Using R-precision as the measure, the reference Boolean run and several of the best ranked runs were equally effective.

While the average R-precision for the better runs was approximately the same, different runs were relatively better for different topics and each run found relevant documents that the other systems did not retrieve. In particular, the collection contains many relevant documents that do not match the negotiated Boolean queries. This is an important finding for current practice since legal discovery is a recall-oriented task.

3.5 The question answering (QA) track

The goal of the question answering track is to develop systems that return actual answers, as opposed to ranked lists of documents, in response to a question. The 2006 track contained two tasks, the main task that was a series task similar to the task used in TRECs 2004 and 2005, and a complex interactive QA (ciQA) task.

The questions in the main task were organized into a set of series. A series consisted of a number of "factoid" (questions with fact-based, short answers) and list questions that each related to a common, given target. The final question in a series was an explicit "Other" question, which systems were to answer by retrieving information pertaining to the target that had not been covered by earlier questions in the series. Answers were required to be supported by a document from the corpus used in the track, the *AQUAINT Corpus of English News Text* (LDC catalog number LDC2002T31, see `www.ldc.upenn.edu`).

In a change from previous years, time-dependent factoid questions were required to be answered with regard to a particular timeframe (as opposed to the timeframe of an arbitrary document containing an answer). For factoid questions phrased in the present tense, the implicit timeframe was the date of the latest AQUAINT document, i.e., the system was required to answer with the most up-to-date information possible. For factoid questions phrased in the past tense, either the question specified the timeframe (*What cruise line attempted to take over NCL in December 1999?*) or the timeframe of the series that included the question was the implied timeframe (for a target of "France wins soccer's World Cup", the question *Who was the coach of the French team?* is to be interpreted as the coach at the time of the World Cup).

The score for a series was computed as a weighted average of the scores for the individual questions that comprised it, and the final score for a run was the mean of the series scores. In a second change from previous years, the weights given to factoid, list, and other questions in the average were equal. This change lessened the importance of factoid questions in the final score.

In absolute terms, the series scores for participating systems have decreased since 2004. This reflects the increasing difficulty—and realism—of the evaluation conditions. In particular, the new requirement for answers to be correct with respect to the date of the latest document in the collection is a significant departure from previous requirements.

The ciQA task was a blend of the TREC 2005 relationship QA task and the TREC 2005 HARD track. The goal of the task was to extend systems' abilities to answer more complex information needs than those covered in the main task and to provide a limited form of interaction with the user in a QA setting.

The questions used in the task contained two parts, a specific question derived from templates of relationship question types, and a narrative that provided more explanation for the specific question. The system response to a question was a ranked list of information "nuggets" supported by AQUAINT documents, where each nugget provides evidence for the relationship in question.

The limited interaction with the user (using the assessor as the surrogate user) was accomplished through

forms as in previous HARD tracks. Participants were allowed to create one HTML-based form per question per run. The form contained a task for the assessor to perform, and assessors were limited to no more than 3 minutes per form. The result of the interaction with a form were returned to the participant, who (presumably) incorporated the results into a new question answering run.

Six groups participated in the ciQA task. In addition, the University of Maryland provided an initial baseline run constructed by retrieving sentences using the Lucene search engine, and a corresponding final baseline run that eliminated those sentences that the assessor marked not relevant during the clarification form interaction. This baseline set was among the best of the runs, excluding a manual run set that was clearly more effective than all other submissions. This is yet another example in TREC 2006 where it has proved difficult to improve on the effectiveness of standard retrieval technology for more specialized tasks.

Thirty-one groups participated in the QA track.

3.6 The spam track

The spam track was first run in TREC 2005. The immediate goal of the track is to evaluate how well systems are able to separate spam and ham (non-spam) when given an email sequence. Since the primary difficulty in performing such an evaluation is getting appropriate corpora, longer term goals of the track are to establish an architecture and common methodology for a network of evaluation corpora that would provide the foundation for additional email filtering and retrieval tasks. Nine groups participated in the TREC 2006 spam track.

The 2006 track included an on-line filtering task as in the 2005 track, plus an enhancement to that task and a new active learning task. For each task the track used a test jig developed for the track that takes an email stream, a set of ham/spam judgments, and a classifier, and runs the classifier on the stream reporting the evaluation results of that run based on the judgments. In the original on-line filtering task, the classifier receives the correct designation for a message as soon as it classifies the message (this represents ideal user feedback). In the delayed feedback extension to the task, the classifier eventually receives the correct designation for each message, but the designation for a given message m may come after some number of intervening messages that must be classified before the feedback for m is received. In the new active learning task, the classifier must determine the designations for the final 10% of an email stream based on learning the correct designations for exactly N messages of its own choosing from the first 90% of the stream (where N was much smaller than 90% of the collection size).

The track used two private email streams and two public email streams. The private streams and one of the public streams were predominately English streams (some spam messages could be in other languages) while the second public stream was predominately Chinese. Participants ran their own filters on the public corpora using the jig and submitted the evaluation output to NIST. For the private corpora, participants submitted their filters to NIST. NIST passed the filters onto the University of Waterloo after stripping all identification of which filters came from which participant. The University of Waterloo used the jig to run the filters on the private corpora and returned the evaluation results to NIST, who then forwarded the evaluation results to the appropriate participant.

The overall results were consistent across the four email streams. Detecting spam is more difficult when given delayed feedback than when immediate feedback is available; the active learning task is even more difficult. Nonetheless, filters are able to detect the vast majority of spam with high accuracy, and there is no indication that this year's (more recent) spam is any harder to detect than earlier spam.

3.7 The terabyte track

The goal of the terabyte track 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 previous TREC collections.

The document collection used in the track was the same collection created for the initial running of the track in TREC 2004: 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 contains approximately 25 million documents and is 426 GB. The collection is distributed by the University of Glasgow, see `http://ir.dcs.gla.ac.uk/test_collections/`.

The track contained three tasks, a classic ad hoc retrieval task, an efficiency task, and a named-page-finding task. Manual runs were strongly encouraged for the ad hoc task since manual runs frequently contribute unique relevant documents to the pools. As part of the inducement for manual runs, an (unspecified) prize was offered to the group that returned the greatest number of unique relevant documents. The efficiency and named page tasks required completely automatic processing only.

Fifty new information-seeking topics were created by NIST assessors for the track. Manual runs used only these 50 topics; automatic runs were required to use the set of 149 topics created for the track from TRECs 2004–2006. Systems returned the top 10,000 documents per topic. In an attempt to overcome the bias toward topic title word documents described in section 2.1.3, pools were created in multiple stages with only the initial stage using traditional pooling. See the terabyte track overview paper for more details.

The more effective automatic ad hoc runs used a variety of retrieval models. Most of these runs used features such as phrases or term proximity factors, and pseudo-relevance feedback was generally put to good use. None of the top eight runs made special use of anchor text, and only one used link analysis in producing the retrieved set.

The efficiency task was designed as a way of comparing the efficiency and scalability of systems given participants all used their own (different) hardware. The "topic" set was a sample of 100,000 queries mined from web search engine logs. To be selected for the query set, the query was required to have a minimum number of hits in the GOV2 collection. The title fields from the ad hoc and named-page tasks' topics were added to this set but were not distinguished in any way. The queries were distributed in four different sets to represent four query streams. Queries in a given stream had to be processed in the order in which they appeared in the stream, but queries from different streams could be interleaved in any manner. Participants ran their systems using the entire query set and returned the top 20 documents per query plus reported the average processing time per query and the total time for all queries. Finally, participants were asked to submit one run using one of three open-source information retrieval systems whose efficiency characteristics are known as a way of normalizing for hardware differences. The queries corresponding to the ad hoc and named-page topics were used to measure the effectiveness of the efficiency runs.

Both effectiveness and efficiency varied greatly across participants. As to be expected, systems could realize effectiveness gains by being less efficient (i.e., a system's most effective run differed from its most efficient run).

Since the document set used in the track is a crawl of a cohesive part of the web, it can support investigations into tasks other than information-seeking search. One of the tasks that had been performed in the web track in earlier years was a named-page finding task, in which the topic statement is a short description of a single page (or very small set of pages), and the goal is for the system to return that page at rank one. The terabyte named page task repeated this task using the GOV2 collection and a set of target topics created

by the participants.

In contrast to the ad hoc task, the more effective named-page finding runs exploited some combination of link structure, anchor text and document structure (for example, giving greater weight to document title words). The most effective named-page run, `indri06Nsdp` from the University of Massachusetts that had a mean reciprocal rank score of 0.512, used all three factors.

Twenty-one groups participated in the terabyte track.

4 The Future

Initial plans for TREC 2007 were formulated during the TREC 2006 conference. All of the 2006 tracks except the terabyte track will continue into 2007; the terabyte track will pause while the feasibility of collecting and using an even larger document set than GOV2 is explored.

TREC 2007 will contain a new track optimistically called the "Million Query" track. While it is unlikely that a test collection with literally 1,000,000 queries will be constructed, the goal of the track is to test the hypothesis that a test collection built from very many, very incompletely judged queries (topics) is a better research tool than a traditional TREC pooled test collection. Both NIST assessors and TREC participants will judge on the order of 50 documents for a query. Queries will be mined from web search engine logs with existing TREC topics (title fields) included as part of the query set. The documents to be judged will be selected from participant submissions according to a particular sampling strategy such as those suggested by Yilmaz and Aslam [11] or Carterette et al. [3]. (Particular strategies will be randomly assigned to queries.) The expectation is that this will allow different sampling strategies to be compared on both the validity of the resulting test collection and the expense of producing the collection.

Acknowledgements

Special thanks to the track coordinators who make the variety of different tasks addressed in TREC possible. The track summaries in section 3 are based on the track overview papers authored by the coordinators.

References

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

[2] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling. In *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 619–620, 2006.

[3] Ben Carterette, James Allan, and Ramesh Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 268–275, 2006.

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

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

[7] 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.

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

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

[10] 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`.

[11] Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 102–111, Arlington, Virginia, November 2006.

[12] 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 2006

Arizona State University	Australian National University & CSIRO
Beijing University of Posts and Telecommunications	Carnegie Mellon University
Case Western Reserve University	Chinese Academy of Sciences (2 groups)
The Chinese University of Hong Kong	City University London
CL Research	Concordia University (2 groups)
Coveo Solutions Inc.	CRM114
Dalhousie University	DaLian University of Technology
Dublin City University	Ecole des Mines de Saint-Etienne
ErasmusMC, TNO, & University of Twente	Fidelis Assis
Fudan University (2 groups)	Harbin Institute of Technology
Humboldt University, Berlin & Strato AG	Hummingbird
IBM Research Haifa	IBM T.J. Watson Research Center
Illinois Institute of Technology	Indiana University
Institute for Infocomm Research	ITC-irst
Jozef Stefan Institute	Kyoto University
Language Computer Corporation (2 groups)	LexiClone Inc.
LowLands Team	Macquarie University
Massey University	Max-Planck Institute for Informatics
Massachusetts Institute of Technology	The MITRE Corp.
National Institute of Informatics	National Library of Medicine
National Security Agency	National Taiwan University
National University of Singapore	NEC Laboratories America, Inc.
Northeastern University	The Open University
Oregon Health & Science University	Peking University
Polytechnic University	Purdue U. & Carnegie Mellon U.
Queen Mary University of London	Queensland University of Technology
Ricoh Software Research Center Beijing	RMIT University
Robert Gordon University	Saarland University
Sabir Research, Inc	Shanghai Jiao Tong University
Stan Tomlinson	State University of New York at Buffalo
Technion - Israel Institute of Technology	Tokyo Institute of Technology
TrulyIntelligent Technologies	Tsinghua University
Tufts University	UCHSC at Fitzsimons
University of Alaska Fairbanks	University of Albany
University of Amsterdam (2 teams)	U. of Arkansas at Little Rock
U. of California, Berkeley	U. of California, Santa Cruz
University of Edinburgh	University of Glasgow
University of Guelph	University of Hannover
University and Hospitals of Geneva	U. of Illinois at Chicago (2 groups)
U. of Illinois at Urbana-Champaign	University of Iowa
U. of Karlsruhe & Carnegie Mellon U.	University of Limerick
U. Maryland Baltimore County & APL, Johns Hopkins U.	University of Maryland
University of Massachusetts	The University of Melbourne
Universit degli Studi di Milano	University of Missouri-Kansas City
Universite de Neuchatel	University of Pisa
University of Pittsburgh	University of Rome "La Sapienza"
University of Sheffield	University of Strathclyde
University of Tokyo	U. of Ulster & Saint Petersburg State U.
University of Washington	University of Waterloo
University of Wisconsin	Weill Medical College of Cornell U.
York University	

Overview of the TREC-2006 Blog Track

Iadh Ounis, Maarten de Rijke, Craig Macdonald, Gilad Mishne, Ian Soboroff*

trecblog-organisers@dcs.gla.ac.uk

1 Introduction

The rise on the Internet of blogging, the creation of journal-like web page logs, has created a highly dynamic subset of the World Wide Web that evolves and responds to real-world events. Indeed, blogs (or weblogs) have recently emerged as a new grassroots publishing medium. The so-called blogosphere (the collection of blogs on the Internet) opens up several new interesting research areas.

Blogs have many interesting features: entries are added in chronological order, sometimes at a high volume. In addition, many blogs are created by their authors, not intended for any sizable audience, but purely as a mechanism for self-expression. Extremely accessible blog software has facilitated the act of blogging to a wide-ranging audience, their blogs reflecting their opinions, philosophies and emotions. Traditional media tends to focus on "heavy-hitting" blogs devoted to politics, punditry and technology. However, there are many different genres of blogs, some written around a specific topic, some covering several, and others talking about personal daily life [3].

The Blog track began this year, with the aim to explore the information seeking behaviour in the blogosphere. For this purpose, a new large-scale test collection, namely the TREC Blog06 collection, has been created. In the first pilot run of the track in 2006, we had two tasks, a main task (opinion retrieval) and an open task. The opinion retrieval task focuses on a specific aspect of blogs: the opinionated nature of many blogs. The second task was introduced to allow participants the opportunity to influence the determination of a suitable second task (for 2007) on other aspects of blogs, such as the temporal/event-related nature of many blogs, or the severity of spam in the blogosphere.

The remainder of this paper is structured as follows. Section 2 provides a short description of the newly created Blog06 test collection. Section 3 describes the opinion task, and provides an overview of the submitted runs of the participants. Section 4 describes the open task and the submitted proposals. We provide concluding remarks in Section 5.

2 Blog06 Test Collection

For the purposes of the TREC Blog track, there was a need to create a test collection of blog data. Such a collection should be a realistic snapshot of the blogosphere, containing enough blogs as to have recognisable properties of the blogosphere, and over a long enough time period that events should be recognisable. In addition, the collection should exhibit other properties of the blogosphere, such as splogs and comments spam. A new collection, called Blog06, was created by the University of Glasgow.

The collection included a selection of "top blogs" provided by Nielsen BuzzMetrics and supplemented by the University of Amsterdam. Moreover, a selection of blogs of genres accessible to the TREC assessors was

*Iadh Ounis and Craig Macdonald are affiliated to the University of Glasgow, UK; Maarten de Rijke and Gilad Mishne are affiliated to the University of Amsterdam, Netherlands; Ian Soboroff is affiliated to NIST, USA.

Quantity	Value
Number of Unique Blogs	100,649
RSS	62%
Atom	38%
First Feed Crawl	06/12/2005
Last Feed Crawl	21/02/2006
Number of Feeds Fetches	753,681
Number of Permalinks	3,215,171
Number of Homepages	324,880
Total Compressed Size	25GB
Total Uncompressed Size	148GB
Feeds (Uncompressed)	38.6GB
Permalinks (Uncompressed)	88.8GB
Homepages (Uncompressed)	20.8GB

Table 1: Details of the Blog06 test collection, and its corresponding statistics.

included, covering topics such as news, sports, politics, health, etc. Finally, given the particular severity of spam in the blogosphere, a selection of assumed spam blogs (splogs) were inserted to ensure that Blog track participants had a realistic research setting.

The University of Glasgow monitored the resulting 100,649 blog feeds over an 11 week period from December 2005 to February 2006. During that time, XML feeds, their corresponding homepages and permalink documents were fetched and saved. The final collection was shipped to the Blog track participants by the University of Glasgow[1]. The number of permalinks documents, used as a retrieval unit in the TREC 2006 Blog track, is over 3.2 million of documents. Table 1 shows the statistics of the final collection. Further information about the TREC Blog06 test collection, how it was created, and some of its interesting features compared to other Blog datasets, can be found in [1].

3 Opinion Retrieval Task

A key feature that distinguishes blog contents from the factual content used in other TREC tasks is their subjective nature. Many blog queries are person names, both celebrities and unknown, and the underlying users information needs seem to be of an opinion, or perspective-finding nature, rather than fact-finding [2]. Incorporating this type of subjectivity in a retrieval context remains a challenge.

3.1 Task

In the TREC 2006 Blog track, the opinion retrieval task involved locating blog posts that express an opinion about a given target. The target can be a "traditional" named entity, e.g. a name of a person, location, or organisation, but also a concept (such as a type of technology), a product name, or an event. The task can be summarised as *What do people think about X*, *X* being a target. The topic of the post was not required to be the same as the target, but an opinion about the target had to be present in the post or one of the comments to the post. For example, for

[1] Further information on obtaining the Blog06 collection can be found at `http://ir.dcs.gla.ac.uk/test_collections/`

the target "skype", here is an excerpt from a relevant, opinionated blog post:[2]

> *Skype 2.0 eats its young*
> The elaborate press release and WSJ review while impressive don't help mask the fact that, Skype is short on new ground breaking ideas. Personalization via avatars and ring-tones... big new idea? Not really. Phil Wolff over on Skype Journal puts it nicely when he writes, "If you've been using Skype, the Beta version of Skype 2.0 for Windows won't give you a new Wow! experience." ...

The following is an excerpt from an unopinionated post:[3]

> *Skype Launches Skype 2.0 Features Skype Video*
> Skype released the beta version of Skype 2.0, the newest version of its software that allows anyone with an Internet connection to make free Internet calls. The software is designed for greater ease of use, integrated video calling, and ...

While no explicit scenario was associated with the opinion retrieval task, it aims to uncover the public sentiment towards a given entity (the "target"), and hence it can naturally be associated with settings such as tracking consumer-generated content, brand monitoring, and, more generally, media analysis.

3.2 Topics

Topics used in the opinion retrieval task follow the familiar title, description, and narrative structure, as used in topics in other TREC test collections. 50 topics were selected by NIST from a donated collection of queries sent to commercial blog search engines over the time period that the Blogs06 collection was being collected. NIST assessors created the topics by selecting queries, and building topics around those queries. In particular, the *title* fields are the literal queries from the donated search query logs file. Based on the title field, an assessor developed an interpretation of what the searcher who originally submitted the query was looking for. The assessor then searched the Blog06 test collection to see if blog posts with relevant opinions appear in the collection. This searching was by no means complete and no relevance judgements from this phase were retained. Finally, the assessor recorded his/her interpretation of the query in the *description* and *narrative* fields. An example of a topic is included in Figure 1.

3.3 Assessment Procedure

Participants could create queries manually or automatically from the 50 provided topics. They were allowed to submit up to five runs, including a compulsory automatic run using the title-only field of the topic. Moreover, the participants were asked to prioritise runs, in order to define which of their runs would be pooled. Participants were also encouraged to submit manual runs, as such runs are valuable for improving the quality of the test collection. Each submitted run consisted of the top 1,000 opinionated documents for each topic. The *retrieval units* were the documents from the permalinks component of the collection, where there is the post and comments related to it. However, participants were free to use any of the other Blog06 collection components for retrieval such as the XML feeds and/or the HTML homepages.

Pools were formed from the submitted runs of the participants. The two highest priority runs per group were pooled to depth 100. The remaining runs were pooled to depth 10.

[2]Permalink `http://gigaom.com/2005/12/01/skype-20-eats-its-young/`
[3]Permalink `http://www.slashphone.com/115/3152.html`

```
<top>
  <num> Number: 871

  <title> cindy sheehan

  <desc> Description:
  What has been the reaction to Cindy Sheehan and the
  demonstrations she has been involved in?

  <narr> Narrative:
  Any favorable or unfavorable opinions of Cindy Sheehan are
  relevant. Reactions to the anti-war demonstrations she has
  organized or participated in are also relevant.
</top>
```

Figure 1: Blog track 2006, opinion retrieval task, topic 871.

NIST organised the assessments for the opinion retrieval task. However, the relevance judgement of a document for a topic was only made by one assessor, meaning that no assessor disagreement studies could be made. Given a topic and a blog post, assessors were asked to judge the content of the blog post. For the assessment, the *content* of a blog post is defined as the content of the post itself and the contents of all comments to the post. If the relevant content is in a comment, then the permalink is declared to be relevant. Assessments had two levels. The following scale was used for the assessment:

-1 *Not judged.* The content of the post was not examined due to offensive URL or header (such documents do exist in the collection due to spam). Although the content itself was not assessed, it is very likely, given the offensive header, that the post is irrelevant.

0 *Not relevant.* The post and its comments were examined, and does not contain any information about the target, or refers to it only in passing.

1 *Relevant.* The post or its comments contain information about the target, but do not express an opinion towards it. To be assessed as "Relevant", the information given about the target should be substantial enough to be included in a report compiled about this entity.

If the post or its comments are not only on target, but also contain an explicit expression of opinion or sentiment about the target, showing some personal attitude of the writer(s), then the document had to be judged using the three labels below:

2 Contains an explicit expression of opinion or sentiment about the target, showing some personal attitude of the writer(s), and the opinion expressed is explicitly negative about, or against, the target.

3 Same as (2), but contains both positive and negative opinions.

4 Same as (2), but the opinion expressed is explicitly positive about, or supporting, the target.

Posts that are opinionated, but for which the opinion expressed is ambiguous, mixed, or unclear, were judged simply as "mixed" (3 in the scale).

	Opinion-finding MAP	Topic-relevance MAP
Best	0.3004	0.4219
Median	0.1059	0.1699
Worst	0.0000	2.6e-05

Table 2: Best, median and worst MAP measures for the 57 submitted runs.

Group	Run	MAP	R-prec	bPref	P@10
Univ. of Illinois at Chicago	uicst	**0.1885**	**0.2771**	**0.2693**	**0.5120**
Indiana Univ.	woqs2	0.1872	0.2562	0.2606	0.4340
Tsinghua Univ.	THUBLOGMF	0.1798	0.2647	0.2563	0.3600
Univ. of Amsterdam	UAmsB06All	0.1795	**0.2771**	0.2625	0.4640
CMU (Callan)	blog06r2	0.1576	0.2455	0.2458	0.3580
Univ. of California, Santa Cruz	ucscauto	0.1549	0.2355	0.2264	0.4380
Univ. of Maryland	ParTitDef	0.1547	0.2106	0.2256	0.3360
Univ. of Maryland B.C	UABas11	0.0764	0.1307	0.1202	0.2140
Univ. of Arkansas at Little Rock	UALR06a260r2	0.0715	0.1393	0.1357	0.3320
Univ. of Pisa	pisaBlTit	0.0700	0.1502	0.1535	0.2880
Chinese Academy of Sciences	IIIS	0.0621	0.1134	0.1553	0.2000
National Institute of Informatics	NII1	0.0466	0.1030	0.0851	0.3140
Robert Gordon Univ.	rguOPN	0.0000	0.0004	0.0003	0.0000

Table 3: Opinion retrieval results: the automatic title-only run from each of 13 groups with the best MAP, sorted by MAP. Note that 1 group (Fudan Univ.) did not submit a title-only run. The best in each column is highlighted.

A workable definition of *subjective* or *opinionated* content was proposed. In particular, a post has a subjective content if it contains an *explicit expression of opinion or sentiment about the target, showing a personal attitude of the writer*. Rather than attempting to provide a formal definition, the assessors were given a number of examples, which illustrated the various evaluation labels above.

3.4 Overview of Results

Overall, 14 groups took part in the opinion retrieval task. There were 57 submitted runs, including 53 automatic runs, and 4 manual runs. Each group was asked to submit a compulsory automatic title-only run, for comparison purposes. Of the 57 submitted runs, 27 were pooled to depth 100, and the rest to depth 10.

The metrics used for the opinion retrieval task are mean average precision (MAP), R-Precision (R-Prec), binary Preference (bPref), and Precision at 10 documents (P@10). Since the opinion retrieval task is an adhoc-like retrieval task, the primary measure for evaluating the retrieval performance of the participating groups is the MAP. Table 2 shows the average best, median and worst MAP measures for each topic, across all submitted 57 runs. While these are not "real" runs, they provide a summary of how well the spread of participating systems is performing. Table 3 shows the best-scoring opinion-finding title-only automatic run for each group in terms of MAP, and sorted in decreasing order. R-Prec, bPref and P@10 measures are also shown.

Table 4 shows the best opinion-finding run from each group, in terms of MAP, regardless of the topic length used. Interestingly, none of the manual runs submitted by the participating groups were beneficial to their retrieval performance.

Group	Run	Topics	MAP	R-prec	bPref	P@10
Indiana Univ.	woqln2	TDN	**0.2052**	0.2881	**0.2934**	0.4680
Indiana Univ.	wxoqf2	TDN	0.2019	**0.2934**	0.2824	0.4500
Univ. of Maryland	ParTiDesDmt2	TD	0.1887	0.2421	0.2573	0.3780
Univ. of Illinois at Chicago	uicst	T	0.1885	0.2771	0.2693	**0.5120**
Tsinghua Univ.	THUBLOGMF	T	0.1798	0.2647	0.2563	0.3600
Univ. of Amsterdam	UAmsB06All	T	0.1795	0.2771	0.2625	0.4640
CMU (Callan)	blog06r2	T	0.1576	0.2455	0.2458	0.3580
Univ. of California, Santa Cruz	ucscauto	T	0.1549	0.2355	0.2264	0.4380
Fudan Univ.	mcwil2knl	TDN	0.1179	0.1860	0.1920	0.2940
Univ. of Pisa	pisaBlDes	TD	0.0873	0.1765	0.1620	0.3400
Univ. of Maryland B.C.	UABas11	T	0.0764	0.1307	0.1202	0.2140
Univ. of Arkansas at Little Rock	UALR06a260r2	T	0.0715	0.1393	0.1357	0.3320
Chinese Academy of Sciences	IIIS	T	0.0621	0.1134	0.1553	0.2000
National Institute of Informatics	NII1	T	0.0466	0.1030	0.0851	0.3140
Robert Gordon Univ.	rguOPN	T	0.0000	0.0004	0.0003	0.0000

Table 4: Opinion retrieval results: one run from each of 14 groups with the best MAP, sorted by MAP. Note that all runs in this table were automatic. The best in each column is highlighted. (An extra row was added to show the run (wxoqf2) with the highest R-Prec). T, TD & TDN respectively denote whether the title field, the title and description fields, or the title, description and narrative fields of the topic files were used by the particpant for that run.

In the qrels provided by NIST, documents with adhoc-style relevance to the query (judged 1 or above, as described in Section 3.3 above) were also included. This allows evaluation of the submitted runs based on the relevance of their returned documents. Table 5 reports the best run from each group in terms of topic-relevance. Moreover, Table 6 reports the Spearman's ρ and Kendall's τ correlation coefficients between opinion-finding and topic-relevance measures. The overall rankings of systems on both opinion-finding and topic-relevance measures are extremely similar, as stressed by the obtained high correlations. Figure 2(a) shows a scatter plot of opinion-finding MAP against topic-relevance MAP, which confirms that the correlation is very high.

For the 57 submitted runs, Figure 2(b) plots both opinion-finding MAP and topic-relevance MAP, sorted by opinion-finding MAP. Noticeable from this plot is that runs appear to be clustered into two groups, those above 13% opinion-finding MAP, and those below (see also Table 4). It is interesting that even runs with medium topic-relevance performance can still do comparatively well on opinion-finding MAP compared to runs with stronger topic-relevance performance. In particular, run *uicst* from Univ. of Illinois at Chicago is noticeable as being a strongly performing opinion-finding run compared to its topic-relevance performance.

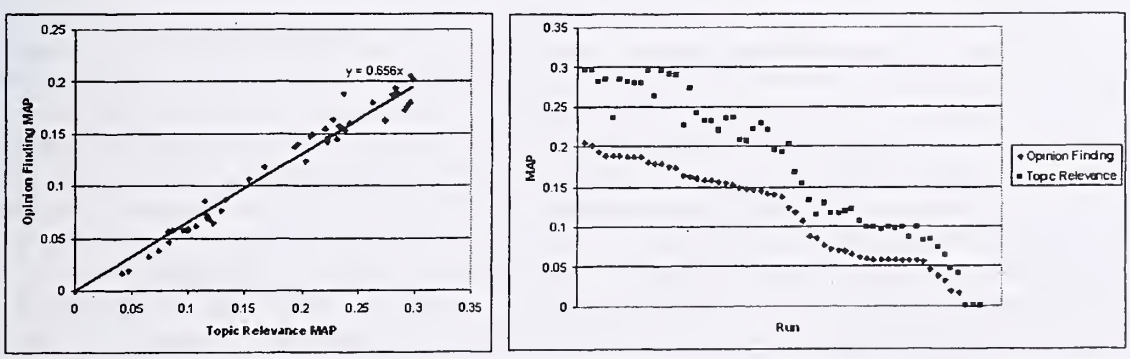
If we rank all the submitted 57 runs by MAP (see Figure 2(b)), for the opinion-finding task, we can determine how many of the top runs are not statistically different, using the Wilcoxon rank test. In particular, of all 57 runs from the opinion-finding MAP, the 9 runs from the best run until run id wxoqs2 (MAP 0.1798) have no significant difference from the best run (woqln2, MAP 0.2052). For topic-relevance MAP, there are some marked differences between the performing systems. In the top 15 runs, 7 are statistically different to the top run, while 8 are not.

Group	Run	Topics	MAP	R-Prec	bPref	P@10
Indiana Univ.	wxoqf2	TDN	**0.2983**	**0.3925**	0.4225	0.6500
Indiana Univ.	woqln2	TDN	0.2963	0.3892	**0.4272**	0.6720
Tsinghua Univ.	THUBLOGMF	T	0.2959	0.3816	0.4177	0.6080
Univ. of Maryland	ParTitDesDef	TD	0.2849	0.3490	0.3998	0.6200
Univ. of Amsterdam	UAmsB06All	T	0.263	0.3674	0.3849	0.6940
Univ. of Illinois at Chicago	uicst	T	0.237	0.3315	0.3415	0.6860
CMU (Callan)	blog06r2	T	0.2324	0.3470	0.3599	0.5480
Univ. of Illinois at Chicago	uicsr	T	0.2267	0.3278	0.3410	**0.7060**
Univ. of California, Santa Cruz	ucscauto	T	0.2203	0.3047	0.3312	0.6480
Fudan Univ.	mcwil2knl	TDN	0.1668	0.2589	0.2826	0.4400
Univ. of Pisa	pisaBlDes	TD	0.1327	0.2329	0.2328	0.5880
Univ. of Maryland B.C.	UABas11	T	0.1288	0.1805	0.1911	0.4520
Univ. of Arkansas at Little Rock	UALR06a500r4	T	0.1192	0.1950	0.1966	0.5180
Chinese Academy of Sciences	IIIS	T	0.1071	0.1903	0.2673	0.3400
National Institute of Informatics	NII1	T	0.0834	0.1522	0.1345	0.5640
Robert Gordon Univ.	rguOPN	T	0.0001	0.0010	0.0010	0.0060

Table 5: Topic-relevance results: documents with 1 or above as relevance label as per the relevance scale defined in Section 3.3. One run from each of 14 groups with the best MAP, sorted by MAP. Note that all runs in this table were automatic. The best in each column is highlighted. (Two extra rows were added to show the runs with the best bPref and P@10, woqln2 and uicsr respectively).

Evaluation Measure	ρ	τ
MAP	0.9745	0.8835
R-Prec	0.9649	0.8609
bPref	0.9505	0.8434
P@10	0.9597	0.8521

Table 6: Correlation of system rankings between opinion-finding performance measures and topic-relevance performance measures. Both Spearman's Correlation Coefficient (ρ) and Kendall's Tau (τ) are reported.



(a) Scatter plot of opinion-finding MAP against topic-relevance MAP.

(b) Opinion finding MAP vs topic-relevance MAP, sorted by opinion-finding MAP.

Figure 2: Figures examining opinion-finding and topic-relevance MAP.

Relevance Scale	Label	Nbr. of Documents	%
Not Judged	-1	0	0%
Not Relevant	0	47491	70.5%
Adhoc-Relevant	1	8361	12.4%
Negative Opinionated	2	3707	5.5%
Mixed Opinionated	3	3664	5.4%
Positive Opinionated	4	4159	6.2%
(Total)	-	67382	100%

Table 7: Relevance assessments of documents in the pool.

Relevance Scale	Nbr. of Splog Documents
Not Judged	0
Not Relevant	8348
Adhoc-Relevant	1004
Negative Opinionated	191
Mixed Opinionated	160
Positive Opinionated	290
(Total)	9993

Table 8: Occurrences of presumed splog documents in the pool

3.5 Overview of the Relevance Judgements

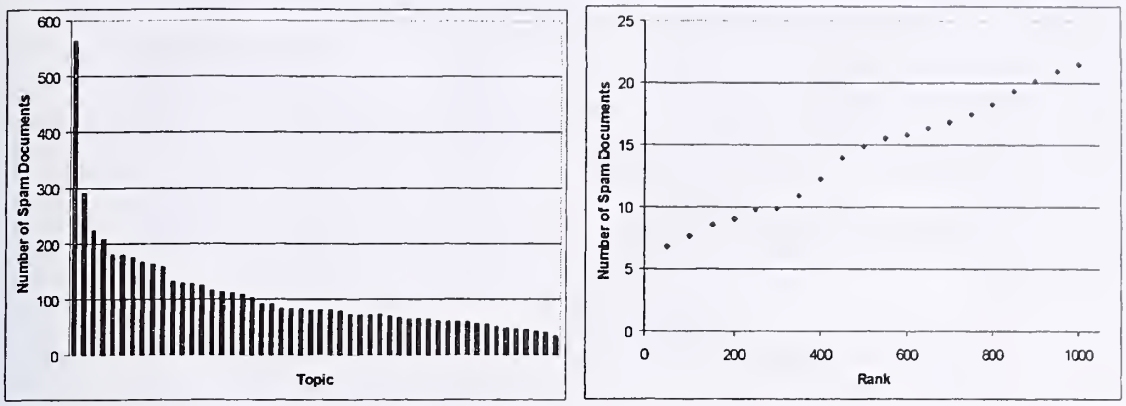
Table 7 shows the breakdown of the relevance assessment of the pooled documents, using the assessment procedure introduced in Section 3.3. About 70% of the pooled documents were judged as irrelevant. As described above, the '-1' element was introduced to allow assessors to discard documents if their associated URL was offensive. However, no assessors made use of this element, choosing in fact to judge all pooled documents. Moreover, it is of note that roughly an equal percentage of opinionated documents were of positive, negative and mixed opinions.

3.5.1 Spam Documents

Since spam is thought to be an issue in the blogosphere, and given that our test collection included a list of assumed splog feeds, we tried to determine the extent to which splog posts had infiltrated the pool, and affected the retrieval systems of the participants. The 17,958 splog feeds in the Blog06 collection generated 509,137 posts. Table 8 provides details on the number of presumed splog posts which infiltrated each element of the relevance scale. In total, 9,993 assumed splog documents were pooled, less than 2% of the splog posts in the collection. Moreover, most assumed splog documents were found not to be opinionated, though those that were were mostly positive.

Figure 3(a) shows the average number of spam documents retrieved by all 57 submitted runs for each topic, in decreasing order. Noticeably, topic 899 (namely "cholesterol") has by far the largest number of splog posts retrieved in the submitted runs (average 564 documents per run). Topic 893 also had a substantial number of splog posts retrieved (average 292 documents per run) - this was again a health topic "zyrtec", which is a medication. Topics which retrieved far fewer spam documents, were concerning people, such as topics 854 and 871 ("Ann Coulter" (34 documents) and "cindy sheehan" (43 documents), respectively).

Next, we examined how the participating systems had been affected by spam documents. Figure 3(b) shows



(a) For each topic, the average number of spam documents retrieved by all of the 57 submitted runs, in decreasing order.

(b) The average number of spam documents retrieved by range of ranks (50), across all topics and submitted runs.

Figure 3: Figures examining the presence of spam documents, by topic and by ranks.

the distribution of spam documents by range of ranks (in units of 50), across all 57 submitted runs and all topics. From this, we can see that on average, systems retrieve more spam documents at later ranks than earlier ranks. In particular, the average number of spam documents retrieved by all systems in the top 10 documents was 1.3. This indicates that the participating systems were good at retrieving non-splog posts at top ranks, and that splog documents were not likely to be retrieved at early ranks. In particular, for the best opinion-finding MAP run of each group, Table 9 shows the mean number of splog documents in the top 10 ranked documents (denoted Spam@10), for all the retrieved documents (Spam@all), and finally BadMAP, which is the Mean Average Precision when the spam documents are treated as the relevant set. BadMAP shows when spam documents are retrieved at early ranks (a low BadMAP value is good, high BadMAP is bad as more spam documents are being retrieved at early ranks). From this table, we can see that some runs were less susceptible to spam documents than others. In particular, runs from the Univ. of Illinois at Chicago and the Univ. of Pisa exhibit the lowest BadMAP values (It is pertinent to note that the Univ. of Pisa reported removing splogs from their collection). In contrast, the run ParTiDesDmt2 of the Univ. of Maryland was affected much more by splog documents.

We also examined the correlation between the official opinion-finding MAP measure calculated using the official relevance assessments (which include spam), and when the assumed spam was removed from the relevance assessments (denoted MAP_NoSpam). Over the 57 submitted runs, the correlation was extremely high ($\rho = 0.9956, \tau = 0.9649$), showing that there is little difference in the overall ranking of submitted runs if the assessors assessed spam or not.

To see if runs that retrieved more spam documents were more likely to be high performing systems or low performing systems, we correlated the ranking of submitted runs by BadMAP, correlating this with MAP_NoSpam. However, the correlation was low ($\rho = 0.2769, \tau = 0.1805$), showing that indeed there was no strong relation between the opinion-finding MAP performance of systems and their likeliness to retrieve spam. However, as the correlation was not negative, it is not the case that low performing systems were more likely to retrieve spam.

Overall, while the Blogs06 test collection contained a component of assumed splogs, the above conclusions suggest that these were not a major hindrance to the retrieval performance of participating groups. Moreover, some topics were more pre-disposed to spam (for example, topics about health), suggesting that these could be identified by statistical predictors.

Group	Run	Spam@10	Spam@all	BadMAP $*10^{-5}$
Indiana Univ.	woqln2	0.78	140.56	6.2
Univ. of Maryland	ParTiDesDmt2	1.92	172.74	14.0
Univ. of Illinois at Chicago	uicst	0.80	38.80	1.0
Tsinghua Univ.	THUBLOGMF	1.56	118.34	5.0
Univ. of Amsterdam	UAmsB06All	0.96	128.22	4.8
CMU (Callan)	blog06r2	0.98	66.02	2.8
Univ. of California, Santa Cruz	ucscauto	1.00	105.38	6.4
Fudan Univ.	mcwil2knl	1.42	60.46	2.4
Univ. of Pisa	pisaBlDes	0.60	48.74	1.6
Univ. of Maryland B.C.	UABas11	1.34	112.74	5.6
Univ. of Arkansas at Little Rock	UALR06a260r2	0.94	95.86	3.0
Chinese Academy of Sciences	IIIS	0.92	75.30	4.2
National Institute of Informatics	NII1	1.44	63.34	4.6
Robert Gordon Univ.	rguOPN	1.56	160.58	6.2

Table 9: Spam measures for runs from Table 4, in the order given. Spam@10 is the mean number of spam posts in the top 10 ranked documents, Spam@all is the mean number of spam posts for each topic. BadMAP is the Mean Average Precision when the spam documents are treated as the relevant set. This shows when spam documents are retrieved at high ranks (low is good, high is bad).

3.5.2 Polarity

We examined the extent to which the submitted runs identified positive and negative opinionated documents. However, because participant systems were not required to rank positively or negatively opinionated documents, the use of precision type measures is not suitable. Therefore, we only look at the recall performance for this analysis. Figure 4 shows the recall of each system in terms of positively opinionated documents against negatively opinionated documents. The gradient of the trend line (0.9342) shows that appears to be a slight overall tendency of the systems to retrieve positively opinionated documents.

Table 10 takes the per-topic best and median runs of the 57 submitted runs, and measures their positive and negative recall. Interestingly, it shows that the best systems are almost equally good at retrieving positive or negative opinions, while the median runs are slightly better at retrieving negatively opinionated documents.

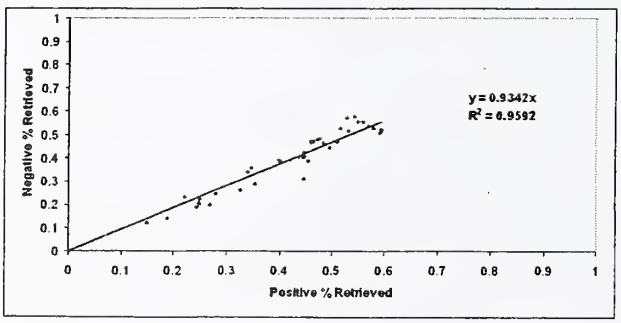


Figure 4: Correlation of positive and negative Recall, by system, over the 57 submitted runs.

	Positive Opinionated Recall	Negative Opinionated Recall
Best Runs	0.7814	0.7754
Median Runs	0.3951	0.4177

Table 10: Recall of positively and negatively opinionated documents, for the per-topic best and median runs.

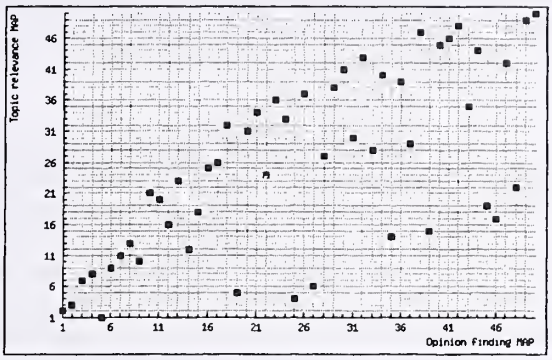


Figure 5: Scatter plot of median opinion-finding MAP against median topical-relevance MAP.

3.5.3 Per Topic Analysis

Analysing further in a per-topic manner allows us to make the following observations. The median number of (topically) relevant documents (i.e., scoring at least 1) per topic was 329, while the median number of documents scoring at least 2 was 182 per topic. The median of the fraction of assessed documents scoring at least 2 vs documents scoring at least 1 was 67% (with minimum 5% and maximum 99%). So overall, for most topics there were more relevant documents than opinionated documents, however, the proportion of opinionated documents varied highly over the topics.

Topics for which median performance (in terms of opinion finding MAP) was high consisted mostly of named entities ("Heineken" (883), "netflix" (863), "Ann Coulter" (854)), whereas low-scoring topics included a mix of such entities along with high-level concepts ("cholesterol" (899), "Business Intelligence Resources" (898)).

As to topics for which the difference between best and median performance was the largest: it seems difficult to define any pattern. These topics are 859, 863, 865, 877, 883, 892 ("letting india into the club," "netflix," "basque," "sonic food industry," "heineken," "jim moran"), and they vary on many dimensions — number of relevant documents, average precision, etc. The same holds true for the topics with the smallest differences between best and median performance: 879, 896, 882, 891, 897 ("hybrid car," "global warming", "seahawks," "intel," "ariel sharon").

Overall, there is a moderate (but not a strong) positive correlation between difficulty in terms of topical relevance and difficulty in terms of opinion finding (Spearman's ρ: 0.6898). Figure 5, which shows a scatter plot of median opinion finding MAP vs median topical MAP (topics sorted by opinion finding MAP), confirms this point. In summary, we infer that the success of the opinion-finding approaches was higher for easier topics.

3.6 Participant Approaches

Looking into the retrieval techniques deployed by the 14 participants, we noticed that most participants approached the opinion retrieval task as a two-stage process. In the first stage, documents are ranked based on topical rele-

vance only, using, mostly, off-the-shelf retrieval systems and weighting models. For example, the University of Maryland, Baltimore County/John Hopkins University (UMBC/JHU) and the Univ. of Arkansas at Little Rock (UALR) used a TF*IDF document ranking scheme. The National Institute of Informatics (NII), the University of Amsterdam (UAmsterdam), the University of Maryland (UMaryland), Chinese Academy of Sciences (CAS), Robert Gordon Univ. and Carnegie Mellon Univ. used language modelling approaches. Finally, the University of Illinois at Chicago (UIllinois) and the University of Pisa (UPisa) used other probabilistic approaches. In the second stage, these results are re-ranked or filtered by applying one or more heuristics for detecting opinions in the documents retrieved at the first stage. The reported approaches by participants for the identification of opinionated content include:

- Dictionary-based approaches: In these approaches, lists of terms and their semantic orientation values were used to rank documents based on the frequency of such words in them, sometimes combined with information about the distance between the sentiment-oriented words and occurrences of query words in the document. In particular, NII proposed a generative language model that models the occurrences of topic terms and opinion-bearing terms in documents. The lists used (e.g., by UMaryland, UIndiana, UAmsterdam, CAS, Tsinghua University (THUIR), UPisa or UMBC/JHU) were either manually-compiled or created automatically. Reports on the success of this approach varied, with some groups observing slight degradation of results compared to their base retrieval scores, and others observing some improvement.

- Text classification approaches: Using training data taken from sources known to contain opinionated content (such as web sites specialising in product reviews) and sources assumed to contain little opinionated content (such as online encyclopedias or news collections), a classifier was trained and used to estimate the degree of opinionated content in retrieved documents. Most groups who used this approach (e.g., UIllinois, UCaliforniaSC, and UALR) favoured Support Vector Machines for their classification, although other classifiers were also used. The success of this approach was limited, possibly because of the difference between training data and the actual opinionated content in blog posts.

- Shallow linguistic approaches: some participants (e.g., UIndiana) used the frequency of pronouns or adjectives as indicators of opinionated content; again, the success of this approach was limited.

In addition to measuring the effect of opinion-detection heuristics, some participants evaluated the benefit of using traditional IR techniques, such as passage retrieval (e.g., UMaryland), or query expansion using pseudo-relevance feedback (e.g., UAmsterdam, UIllinois, or UCaliforniaSC), or using external corpora, (e.g., UIllinois, or UMBC/JHU). Finally, some participants specifically addressed noise in the collection, evaluating the effectiveness of spam detection and other noise removal techniques from the retrieved results (e.g., UIndiana, UPisa, UAmsterdam, or UMBC/JHU). It is difficult to assess the overall effectiveness of these approaches without experimental baselines.

3.7 Baseline Systems

As mentioned above, most participating groups deployed systems using a two-stage process. We desired to assess the usefulness of the post-processing layer in extracting opinionated documents, when compared to a standard IR system. To this end, some additional runs were produced by the organisers, using standard off-the-shelf IR systems, without any opinion-finding specific features. Note that none of these runs were in the assessment pool. Table 11 shows the retrieval performances achieved by the in-house NIST Prise v3, and by the open-source version of Terrier from the University of Glasgow[4]. The indexing settings for Terrier are Porter's stemming and standard

[4]Terrier can be downloaded from `http://ir.dcs.gla.ac.uk/terrier/`

Systems	Topic Fields	Opinion MAP	Relevance MAP
Prise v3	TD	0.1858	0.2908
Terrier v 1.0.2	T	0.1696	0.2703
Terrier v 1.0.2	TD	0.2115	0.3151
Terrier v 1.0.2	TDN	0.1992	0.2892
Terrier v 1.0.2	TN	0.1655	0.2402

Table 11: Performance achieved by standard baseline IR systems. For Terrier, the PL2 weighting model was used with its default parameters.

stopword removal; the DFR PL2 weighting model and its default setting is applied to rank the documents. In particular, it is noticeable that the TD run of Terrier would have achieved the best run on both opinion-finding and topic-relevance MAP measures.

Looking at the 57 submitted runs, there were 42 runs using the title fields of the topics only; 10 using the title and description; and 5 using title, description and narrative. Due to this high variation, it is not possible to draw conclusions as to whether the description and narrative fields helped retrieval for participating systems. However, our baseline runs using Terrier anecdotally suggest that the description field of the topics was beneficial, but the narrative was not.

For future years, participants may benefit from the provision of stronger topic-relevance baseline runs, or detailed instructions on how to use off-the-shelf IR systems, similar to the comparative run systems deployed in the TREC 2006 Terabyte track.

4 Open Task

In the initial proposal for the Blog 2006 track, the intention was to run a time-oriented task, called event timelining, as a second task for TREC 2006. The idea was to focus on the chronological publication order and the associated importance of time in the blogosphere. However, during the TREC 2005 workshop on the Blog track, the workshop participants did not find event timelining to be too interesting, or to be only specific to the blogs. Instead, it was agreed to set up an open task aimed at defining a suitable task for TREC 2007.

Unlike the opinion retrieval task, the open task was not set up as an evaluation task. The open task was meant to provide participants with an opportunity to explore other aspects of blogs besides their opinionated nature. That is, we invited participants to define their own task, which c could be sensibly operationalised and then evaluated in a way that reasonably abstracts the user task. For inspiration, a number of possibilities were sketched in the guidelines for participants, including authority detection (e.g. use part of the corpus to estimate the indegree rank of another part of the corpus), temporal event mining (e.g. identify and follow reactions from bloggers to events which fall under the users' areas of interest), blog finding (e.g. locate blogs about a given topic, rather than posts), spam blog classification, etc.

Participants were asked to propose a "TREC-style" task, which could be used for the TREC 2007 Blog track. This means that the results of the task can be evaluated by a team of assessors and that different approaches can be compared. Groups taking part in the open task were asked to submit a paper describing their proposed task in two steps: First, submit a short abstract, including the definition of the task, some motivation on why it is useful in a realistic blog retrieval environment, and a brief description of the proposed assessment procedure (e.g., how is the task being evaluated?). Secondly, submit a full paper providing a thorough discussion of the proposed task.

4.1 Participants and Results

In total, five proposals were submitted to the open task. They are briefly described below.

NEC Laboratories America proposed the task of identifying spam blogs (splogs) in the collection. Spam blogs are a serious issue in the blogosphere, and their elimination may be a key part in improving results of blog retrieval and other tasks involving blogs. Specifically, the suggested tasks included identification of splogs with fixed training and test sets, and an adaptive splog identification task, where the performance is measured incrementally, as more and more data is available to the system.

The University of Maryland Baltimore County and Johns Hopkins University also proposed the task of splog detection, where the collection is split in time, the first part used for training and the second for testing. Participants would be required to identify splogs in the test collection, and possibly also suggest the type of spamming method being used. An additional extension of the task would evaluate the contribution of spam detection and removal to retrieval performance.

Robert Gordon University proposed a task related to the identification of emerging trends in blogs: topics, which are discussed substantially more during a specific time interval than during preceding intervals. Participants in this task would be given a set of topics and training intervals for each, and would be required to predict those topics that would become "hot topics" during a test interval. Possible approaches to deciding whether a topic is an emerging trend include the volume of discussion about the topic in terms of number of posts or their length, as well as the relation between the topic and other topics.

CSIRO ICT Centre advocated the idea that the availability of more information about a situation and person, i.e. context, will lead to better results for users of search systems. They proposed a sentiment-related task where the blogger's sense-of-self and its changes over time are analysed from the blog posts. In particular, participants in this task would be required to identify, as a first stage, those bloggers who display substantial changes in sense-of-self over time, and, as a second stage, the blog posts which contribute most to tracking these changes. Identifying the blogger's sense-of-self is seen as a partial approach to providing context to the retrieval of blog posts. What this deeper context may add to explicit and implicit search is touched on.

The National Institute of Informatics, Japan proposed a task that is similar to the Story Link Detection at the TDT evaluation, and which involves identifying whether two blog posts discuss the same topic. Participants of this task would be given a set of pairs of blog posts, and would return, for each pair, a decision on whether the two posts are linked — meaning that they share the same topic. Applications of this task include summarisation and "related posts" suggestion.

All the above participants were invited to a (separate) pre-TREC 2006 workshop, to discuss their task proposals. The main purpose of the workshop is to plan the track activities for TREC 2007. While the workshop outcome did not lead to a clear consensus on the submitted proposals, two possible tasks have emerged:

- An information filtering-like task, e.g. *Inform me of new blog entries about X*, *X* being a target as in the opinion retrieval task.
- A Blog expert-like task, e.g. *Find the best blog entries about X, or bloggers with a (recurring) interest in X*, *X* is again a target.

Both above tasks address some interesting features of the blogosphere, and are currently being investigated for a second evaluated task in the TREC 2007 Blog track.

5 Conclusions

TREC 2006 was the first year the Blog track was run. A new large test collection of blog data, called Blog06, was created, and a particular feature of blogs has been tackled, namely the opinionated nature of posts on the blogosphere. The participants results suggest that this task is challenging, and requires further investigation. We found that the retrieval performance on the opinion retrieval task is strongly dominated by the performance on the underlying topic relevance task, emphasising the importance of a strong retrieval baseline. We also found that the pooled documents were not infiltrated by spam to any great extent, and the presence of the spam in the pool did not affect the overall ranking of systems. Moreover, there was no strong evidence that the participating systems retrieved one kind of opinion over another. Finally, there seems to be a positive but not strong correlation between difficulty of topics, in terms of opinion-finding MAP and topic-relevance MAP. It is hoped that by using the relevance assessments of this year as training data, participants will be able to further their techniques for identifying opinionated blog posts.

For the open task, there was no clear emerging task suitable for evaluation in TREC 2007. However, an information filtering or blog identification task (as discussed in Section 4.1) seem to address some interesting elements of the blogosphere, and could be run in forthcoming Blog tracks.

Task details for TREC 2006 Blog track are maintained on the track wiki, at `http://www.science.uva.nl/research/iiwiki/wiki/index.php/TREC-blog`.

Details on the TREC 2007 Blog track are provided on the following wiki page: `http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG`

Acknowledgements

We are grateful to Nielsen BuzzMetrics/BlogPulse for contributing a feed list for creating the Blog06 test collection. We also would like to thank Matthew Hurst and Janyce Wiebe for useful advice and discussions on setting up the opinion retrieval task. We would like to thank Ali Azimi Bolourian for monitoring some of the crawls during the Blog06 test collection creation. Finally, we would like to thank Gianni Amati for various comments on the Blog track.

Gilad Mishne and Maarten de Rijke were supported by the Netherlands Organization for Scientific Research (NWO) under project number 220-80-001.

References

[1] Craig Macdonald and Iadh Ounis. The TREC Blog06 Collection : Creating and Analysing a Blog Test Collection *DCS Technical Report TR-2006-224*. Department of Computing Science, University of Glasgow. 2006. `http://www.dcs.gla.ac.uk/~craigm/publications/macdonald06creating.pdf`

[2] Gilad Mishne and Maarten de Rijke. A Study of Blog Search. In *Proceedings of ECIR-2006*. LNCS vol 3936. Springer 2006.

[3] Amanda Lenhart and Susannah Fox. Bloggers : a portrait of the Internet's new storytellers *Pew Internet & American Life Project*. July. 2006.

Overview of the TREC 2006 Enterprise Track

Ian Soboroff
NIST, USA
ian.soboroff@nist.gov

Arjen P. de Vries
CWI, The Netherlands
arjen@acm.org

Nick Craswell
MSR Cambridge, UK
nickcr@microsoft.com

1 Introduction

The goal of the enterprise track is to conduct experiments with enterprise data — intranet pages, email archives, document repositories — that reflect the experiences of users in real organizations, such that for example, an email ranking technique that is effective here would be a good choice for deployment in a real multi-user email search application. This involves both understanding user needs in enterprise search and development of appropriate IR techniques.

The enterprise track began in TREC 2005 as the successor to the web track, and this is reflected in the tasks and measures. While the track takes much of its inspiration from the web track, the foci are on search at the enterprise scale, incorporating non-web data and discovering relationships between entities in the organization. As a result, we have created the first test collections for multi-user email search and expert finding.

This year the track has continued using the W3C collection, a crawl of the publicly available web of the World Wide Web Consortium performed in June 2004. This collection contains not only web pages but numerous mailing lists, technical documents and other kinds of data that represent the day-to-day operation of the W3C. Details of the collection may be found in the 2005 track overview (Craswell et al., 2005). Additionally, this year we began creating a repository of information derived from the collection by participants. This data is hosted alongside the W3C collection at NIST.

There were two tasks this year, email discussion search and expert search, and both represent refinements of the tasks initially done in 2005. NIST developed topics and relevance judgments for the email discussion search task this year. For expert search, rather than relying on found data as last year, the track participants created the topics and relevance judgments. Twenty-five groups took part across the two tasks.

2 Email discussion search task

This task focuses on searching the `lists` subcollection, which are 198,394 pages crawled from `lists.w3.org`, the archive of the W3C mailing lists. Each page contains either a single email or a monthly listing. The messages are rendered into HTML, so participants can treat it as a web/text search or they can recover the email structure (threads, dates, authors, lists) and incorporate this information in the ranking.

One can imagine many different kinds of searches in a mailing list archive. We have focused on searching for discussions and arguments about design and development issues within the W3C.

pop-up ads rely upon javascript to "pop up" OnLoad - that is, when the requested document is parsed by the user agent... since the "pop up" is part of the user interface, if a site employing pop-up ads claims conformance to WCAG, then the markup employed in pop-up adds are also subject to WCAG, while control over the popping is addressed by the User Agent Accessibility Guidelines (UAAG)

no matter the source of the content that pops up, if the site which utilizes pop-up ads does not ensure that the pop ups are WCAG compliant, then that site, or the document to which the OnLoad event that causes a new viewport to be generated is attached (if the claim is document-specific) cannot be considered WCAG compliant... for starters, pop-up ads are not rendered by non-javascript-aware browsers, such as lynx, which means that some users do not have access to all of the content on the page/site – regardless of whether that content is useful...

moreover, as david p has pointed out, turning off scripting in order to suppress the generation of pop-up ads is far too draconian a solution –

Figure 1: Part of an email arguing against the usability of pop-up ads. Note that the topic (DS64) is about pop-up ads, software to block them, and their relative advantages and disadvantages.

Over the course of their standards work, many decisions are made, sometimes after considerable and perhaps contentious debate. In the discussion search task, the goal of systems is to find those discussions, and in particular those messages where different sides of the debate are argued.

2.1 Topics and relevance judgments

In the first year of the track, the topics and relevance judgments for the discussion search task were created by the participants. This was not only due to limited resources at NIST, but primarily because it was thought that the technical nature of the collection was not well-matched to NIST assessor expertise. The experience of developing the collection within the community led us to reconsider this assumption, and so this year NIST assessors developed the topics and made the relevance judgments.

NIST developed fifty topics each of which describe a subject of discussion on the W3C mailing lists. These topics range from differences in the P3P 1.0 and 1.1 recommendations to blocking pop-up windows to evaluating color contrast for color-blind users. Participants were to search for on-topic emails that contain a pro or con argument. For example, a message relevant to the pop-up blocking topic with a negative argument is shown in Figure 1.

An important part of this task is developing an understanding of the kinds of searches that people would like to make in this collection. Wu et al. arranged last year's topics into several general categories, and observed that some categories were more amenable to pro/con discussion, and also that some categories had better inter-assessor agreement (Wu et al., 2006). This year, the assessors followed Wu's categories in designing their topics, and tried to ensure that pro/con discussion existed for that topic in the collection.

In addition to judging whether a message was irrelevant, on topic, or contained a pro/con argument, we also asked the assessors to try to note specifically whether the message was pro or con. Sentiment and relevance are denoted in the relevance judgments according to the following scale:

Run	MAP	R-prec	bpref	P@5	P@10	P@20	MRR
THUDSTHDPFSM	**0.2858**	**0.3186**	**0.3007**	0.4261	0.4022	0.3674	0.6415
srcbds5	0.2852	0.3179	0.2979	**0.4478**	**0.4370**	**0.3913**	0.6323
DUTDS3	0.2808	0.3110	0.2958	0.4304	0.4022	0.3522	**0.6483**
UAmsPOSBase	0.2590	0.3054	0.2743	0.4174	0.3826	0.3435	0.6028
york06ed03	0.2482	0.3141	0.2838	0.4348	0.3978	0.3620	0.5900
UMaTDMixThr	0.2316	0.2824	0.2539	0.3609	0.3478	0.3413	0.5051
IIISRUN	0.2269	0.2720	0.2442	0.3609	0.3217	0.3152	0.5328
IBM06JAQ	0.2030	0.2481	0.2337	0.3826	0.3391	0.3315	0.5992
uwTsubj	0.1891	0.2404	0.2136	0.3043	0.2913	0.2696	0.4285
InsunEnt06	0.1223	0.2004	0.1543	0.3304	0.3000	0.2652	0.5391

Table 1: Discussion search results for the run with the highest MAP from each group. Scores are computed where judging levels '2' (contains a pro/con) and above are considered relevant. The best score for each measure is highlighted. DUTDS3 is a manual run.

0: not relevant.
1: relevant, does not contain a pro/con argument.
2: relevant, contains a negative (con) argument.
3: relevant, contains both pro and con arguments.
4: relevant, contains a positive (pro) argument.

A 10% random sample of each topic's pool was drawn and given to a second assessor in order to measure agreement. Agreement within the sample was similar to levels found in last years relevance judgments as reported in (Wu et al., 2006). When judgments were thresholded so we could measure agreement on whether a message was relevant at all or not, we find a Cohen's kappa of 0.4. Agreement on whether a message was pro/con as opposed to relevant or nonrelevant was 0.35. The sample was not large enough to measure agreement on pro or con messages alone. Relevance judgments for retrieval tasks tend to have a kappa of around 0.4 (varying somewhat between collections and assessor groups), so these values while low are not unusual.

2.2 Results

Runs were evaluated on retrieval of messages containing a pro/con sentiment (levels 2 and above) as well as just retrieving relevant messages (levels 1 and above). Table 1 shows the top run from each group according to mean average precision in retrieving pro/con messages. Table 2 shows the top run from each group for topic relevance retrieval.

Figure 2 compares the MAP scores between the two rankings. Overall, the two rankings of the runs are very similar, with a Kendall's tau of 0.9 for MAP. Three runs, DUTDS3, york06ed02, and IBM06JAQ, are more highly ranked at pro/con retrieval than they are at relevant message retrieval. DUTDS3, a manual run (i.e., a person was involved in some stage of the query processing), is the eleventh-highest ranked run by MAP on relevant messages, but the third-highest ranked by MAP on pro/con messages.

The strong tau correlation indicates although the runs are trying to focus on pro/con messages, topic relevance is the dominant factor in their document rankings. We further tried to determine if the relative ranking of pro/con and relevant messages was better or worse than random. For each topic in each run, we removed the nonrelevant retrieved documents, and computed the average precision of the residual ranking with only pro/con documents considered relevant. We call this "pro/con AP", and it equals 1 when all pro/con messages are ranked

Run	MAP	R-prec	bpref	P@5	P@10	P@20	MRR
THUDSTHDPFSM	**0.4083**	0.4204	**0.4264**	0.6520	0.6120	0.5590	0.7702
srcbds5	0.4065	**0.4275**	0.4222	0.6520	0.6100	0.5610	0.7917
DUTDS4	0.3891	0.4048	0.4062	0.6000	0.5780	0.5310	0.7004
york06ed03	0.3782	0.4195	0.4128	**0.6840**	**0.6180**	**0.5690**	**0.8048**
UAmsPOSBase	0.3750	0.3991	0.3943	0.6280	0.5920	0.5350	0.7776
UMaTDMixThr	0.3631	0.3963	0.3863	0.5880	0.5820	0.5470	0.7134
IIISRUN	0.3430	0.3769	0.3678	0.6080	0.5640	0.5130	0.7283
IBM06JILAPQD	0.3310	0.3717	0.3709	0.5800	0.5640	0.5040	0.7677
uwTsubj	0.2927	0.3377	0.3112	0.5000	0.4980	0.4590	0.5819
InsunEnt06	0.1872	0.2612	0.2125	0.4720	0.4520	0.4210	0.7085

Table 2: Discussion search results for the run with the highest MAP from each group. Scores are computed where judging levels '1' (relevant to the topic) and above are considered relevant. The best score for each measure is highlighted.

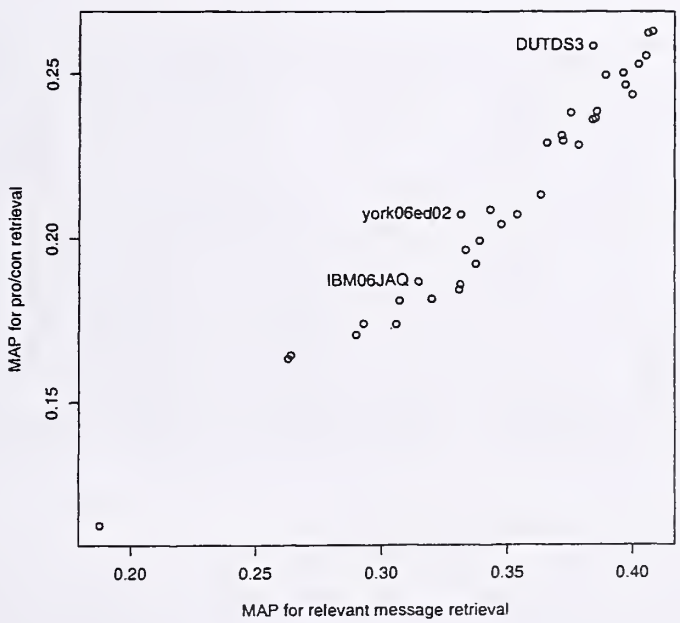


Figure 2: Scatterplot of MAP scores for pro/con and relevant message retrieval. The three labeled runs are ranked more highly for pro/con retrieval than for relevant message retrieval.

ahead of "just relevant" ones. We then generated 1000 random permutations of those pro/con and relevant documents, computing the pro/con AP of each permutation. Sorting the pro/con APs and noting the top and bottom 25 gives a 95% confidence interval on pro/con AP for that number of pro/con and relevant documents. If the actual pro/con AP is above the confidence interval, we would conclude that the run is significantly ordering pro/con documents ahead of relevant ones. Likewise, if the actual pro/con AP is below the interval, we would conclude that the ordering was worse than would be achieved by random shuffling.

Figure 3 presents these results both for each run and for each topic. The top graph shows the number of topics for each run that the actual pro/con AP was above, within, or below the 95% interval. The bottom shows for each topic the number of runs for which their actual pro/con AP was above, within, or below the interval. In each graph, the bar is divided into three sections: the top part counts the topics (or runs) where actual pro/con AP was above the interval, the middle those within the interval, and the bottom those below it. These graphs seem to indicate that most runs do not significantly differentiate relevant and pro/con messages for the majority of topics. Some topics are "easier" in this regard than others, but some are much much harder; note topic 62, where all runs actually ranked the relevant documents ahead of the pro/con ones.

We lastly compared the system ranking for relevant message retrieval to one based on the second assessor's relevance judgments. Since the secondary judgments are only a random sample, we used Yilmaz and Aslam's inferred average precision (infAP) (Yilmaz and Aslam, 2006) measure to estimate average precision for the runs using the sampled judgments. The Kendall's tau correlation of the official MAP ranking to the infAP ranking is 0.695. This is about the same as we saw in last year's judgments, when you consider that the use of a subsample also causes the correlation to be lower. Along with the similarity in agreement measures noted above, this indicates that assessor disagreement for this task is not very different and has about the same effect whether participants or NIST assessors are assessing relevance. We surmise that the lack of familiarity with the W3C and the collection affects all assessor groups strongly.

3 Expert search task

The expert search task is quite different from the traditional TREC search task, in that the goal of the search is to create a ranking of people who are experts in the given topic, rather than relevant documents about the topic. Nick Craswell extracted a canonical list of people addressed in email or on a web page in the W3C collection; this is called the *candidate list.* In response to a given topic, systems return a ranking of candidate experts. In contrast to the email search task, participants may make use of the entire W3C collection. Candidates are pooled and judged for expertise, and the systems are scored using traditional ranked retrieval measures.

The expert search task was the more popular in the track, with 23 groups contributing topics, runs, and/or relevance judgments. There were 91 runs submitted.

3.1 Topics and relevance judgments

In 2005, the enterprise track ran a pilot expert search task where the topics were W3C working groups, and systems were to identify who was part of each working group. The working group truth data came from an official listing of groups and members which was not part of the collection (although some groups were able to find the list by searching the live web). This year, we decided to develop topics for expert search from scratch.

As was done last year for email discussion search, the topics for the expert search task were created and judged by track participants. Twenty groups agreed to help, and each contributed 3-6 topics. Of these, we selected 55 topics for the final set. Once runs were submitted, NIST formed pools and sent them to CWI, where the assessment system was hosted. The topic authors

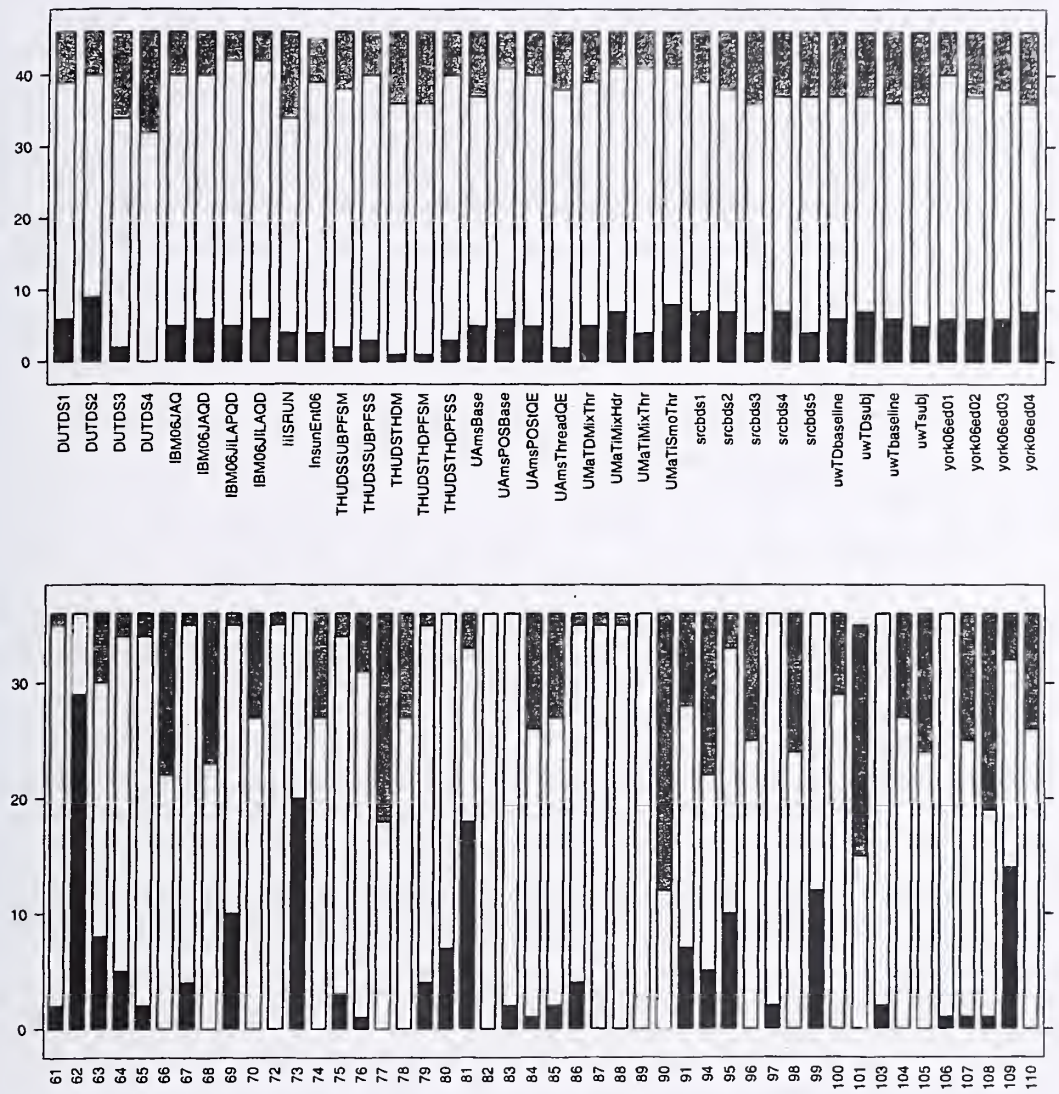


Figure 3: (Top) For each run, the number of topics where actual pro/con AP was better, equivalent to, or worse than random. (Bottom) For each topic, the number of runs whose actual pro/con AP in that topic was better, equivalent to, or worse than random.

Beijing University of Posts & Telecom.	Queen Mary University of London
California State University, San Marcos	Queensland University of Technology
Case Western Reserve University	Robert Gordon University
City University	Shanghai Jiao Tong University
DaLian University of Technology	Tsinghua University
Fudan University	University Amsterdam
University of Glasgow	University of Illinois Urbana-Champaign
IBM	University of Massachusetts
Lowlands Team	University of Waterloo
Open University	University Ulster and
University of Pittsburgh	St. Petersburg State University

Table 3: Groups contributing topics and judgments for the expert search task.

```
<top>
<num> Number: EX51
<title> relationship cardinalities </title>

<desc> Description:
A relevant expert will have knowledge in relationship cardinalities between
roles in different choreographies.
</desc>

<narr> Narrative:
In the context of semantic web, the relationships between entities can have
different cardinalities and roles. Relevant expert will have an explicit knowl-
edge of such choreographies. Experts in Semantic web are not relevant with-
out explicit knowledge in choreographies.
</narr>
</top>
```

Figure 4: A sample expert search task topic.

then judged the pools through the CWI system. We received judgments for 49 of the 55 topics. The names of the groups who contributed their considerable time and effort to this task are listed in Table 3.

A sample topic is shown in Figure 4. Note that this topic resembles a TREC ad hoc topic, except that the user is looking for people rather than documents. The topic statements were composed by the contributor, and only lightly edited to correct the spelling of key words and any ambiguous grammar.

Systems produced a ranked list of expert candidates for each topic. In addition, for each candidate, systems returned a (possibly zero-length) ranked list of documents supporting the designation of that person as an expert in the topic. The purpose of requiring support documents is twofold. First, in an actual application, it is important for the system be able to illustrate why a person is being recommended as an expert. Second, the groups making the relevance judgments could make use of the support documents in deciding whether a person was an expert, rather than doing their own research or relying on background knowledge.

The pools for expert search included the top 20 ranked people for each topic, along with the top 10 support documents for each of those people, from the two highest-priority runs per group. This created very large pools with 6,217 expert-document pairs per topic on average. Ideally,

all support documents are assessed before making a judgment on the candidate's expertise. However, considering the size of the pools, we decided to distinguish between *judged* and *partially judged* expert search topics. In a partially judged topic, the assessment of the candidate's expertise has not been done on the basis of judging all support documents, but using a handful of (positive or negative) support documents only (i.e., some of the pooled support documents are skipped). Unfortunately, making partial judgments did not reduce the workload very much - on average, assessors who judged expertise using partial judgments still made an assessment for more than one out of six support documents in the pool. We explore some possible ideas for reducing the judging load below in the discussion of results.

The relevance scales for expert search are somewhat unusual, to allow for the possibility of indeterminate expertise and support documents which in fact did not support a judgment of expertise either way. The scales used in the expert search relevance judgments were

- Candidate experts:
  - 0: candidate is not an expert.
  - 1: unknown.
  - 2: candidate is an expert.
- Support documents:
  - 0: negative support (document indicates person is not an expert).
  - 1: no support either way.
  - 2: positive support (document indicates person is an expert).

Note that the threshold for correctness for both people and documents is 2, rather than the usual value of 1.

3.2 Results

The evaluation results measure the quality of the ranked list of people using traditional retrieval measures including MAP and precision at fixed ranks. Two sets of measures were provided. The first measures the ranked expert list without regards to the support documents; if a correct expert is returned, the system is credited with returning that expert even if no supporting documents were retrieved. These results are shown in Table 4. Manual runs, where a person was involved at some point in the query process, are shown in italics.

The evaluation scores in Table 5 only gave credit for retrieving a relevant expert if a supporting document was retrieved as well. Credit was awarded if a supporting document appeared anywhere in the list of (maximum) 20 support documents for that person. If no supporting document was retrieved, the person was considered not relevant.

Figure 5 plots each run's no-support-required MAP score against its supported-experts MAP score. The tau correlation of the two rankings is only 0.76. Three runs from ICT did not return any support documents at all, and as such they are found along the x-axis in Figure 5; when we remove those runs from both rankings, the tau improves to 0.82. This is still low enough to indicate noticeable differences in the two rankings. The graph seems to show groups of runs with very closely-scoring expert rankings that differ in their supported-expert ranking. We need to look more closely at those runs that are returning unsupported experts to understand what is happening here more fully.

Run	MAP	R-prec	bpref	P@5	P@10	P@20	MRR
kmiZhu1	**0.6431**	**0.6242**	**0.6391**	**0.8245**	**0.7347**	**0.6031**	**0.9609**
SJTU04	0.5947	0.5783	0.5913	0.7673	0.7041	0.5694	0.9358
SRCBEX5	0.5639	0.5599	0.5642	0.7224	0.6551	0.5469	0.9043
PRISEXB	0.5564	0.5808	0.5614	0.7592	0.6653	0.5459	0.8486
IBM06MA	0.5235	0.5192	0.5180	0.7673	0.6449	0.4857	0.9286
UMaTDFb	0.5016	0.5108	0.5049	0.7265	0.6388	0.5000	0.8571
THUPDDSNEMS	0.4954	0.4978	0.4916	0.6694	0.5939	0.5071	0.8265
ICTCSXRUN01	0.4949	0.4977	0.4858	0.6898	0.5837	0.4908	0.8194
FDUSO	0.4814	0.4989	0.4936	0.7020	0.6306	0.5153	0.8612
UvAprofiling	0.4664	0.4957	0.4707	0.6612	0.5878	0.4959	0.8510
DUTEX2	0.3779	0.4175	0.4077	0.6245	0.5184	0.4184	0.8094
qutmoreterms	0.3673	0.4043	0.3907	0.6327	0.5388	0.4367	0.7683
UMDemailTLNR	0.3503	0.3775	0.3552	0.5388	0.5041	0.4245	0.7064
UIUCe2	0.3364	0.3580	0.3388	0.5388	0.4816	0.3959	0.7187
ex3512	0.3158	0.3425	0.3299	0.5347	0.4612	0.3898	0.7912
uwXSOUT	0.3132	0.3780	0.3364	0.5796	0.5143	0.4112	0.7140
uogX06csnQE	0.3024	0.3433	0.3292	0.5306	0.4429	0.3531	0.7831
PITTPHFREQ	0.2770	0.3513	0.3166	0.5510	0.5041	0.3857	0.7366
sophiarun1	0.2248	0.2864	0.2565	0.4980	0.4306	0.3286	0.6307
w1r1s1	0.2154	0.2818	0.2523	0.5184	0.4245	0.3265	0.6368
l3s2	0.1313	0.1480	0.1401	0.5714	0.2918	0.1459	0.8010
quotes	0.1308	0.1778	0.1844	0.3184	0.2653	0.2224	0.5095
SPlog	0.1126	0.1555	0.1671	0.2531	0.2204	0.1878	0.4709

Table 4: Expert ranking scores without taking support documents into account. The best run in each group according to MAP is shown. Runs in italics are manual runs.

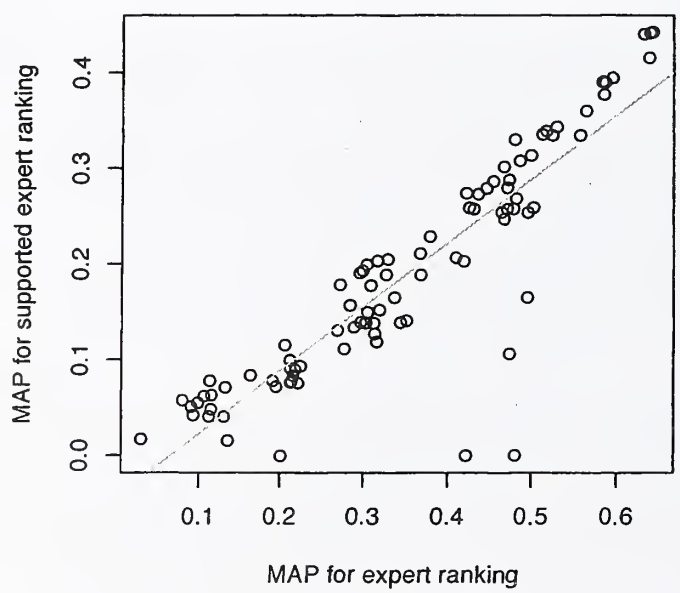


Figure 5: Scatterplot of MAP scores when support is or is not required when considering whether a retrieved person is relevant.

Run	MAP	R-prec	bpref	P@5	P@10	P@20	MRR
kmiZhu1	**0.4421**	**0.4835**	**0.4986**	**0.6612**	**0.5633**	**0.4459**	**0.8369**
SJTU04	0.3943	0.4304	0.4581	0.5714	0.5204	0.4143	0.8132
SRCBEX5	0.3602	0.4092	0.4299	0.5551	0.4735	0.3969	0.7350
IBM06MA	0.3346	0.3829	0.4135	0.5878	0.4878	0.3602	0.7339
PRISEXB	0.3345	0.4203	0.4228	0.5429	0.4571	0.3847	0.6695
UvAprofiling	0.3016	0.3637	0.3743	0.4980	0.4265	0.3582	0.7177
FDUSF	0.2796	0.3148	0.3356	0.4653	0.4041	0.3204	0.6767
UMaTiDm	0.2740	0.3205	0.3350	0.4980	0.4102	0.3204	0.6344
THUPDDFBS	0.2573	0.3035	0.3155	0.4082	0.3673	0.3020	0.6117
DUTEX2	0.2290	0.2918	0.3028	0.4898	0.3898	0.3031	0.6703
qutbaseline	0.2110	0.2561	0.2527	0.4082	0.3531	0.2694	0.6115
ex3512	0.2031	0.2466	0.2724	0.3959	0.3286	0.2786	0.6481
UIUCe2	0.1650	0.2271	0.2582	0.3143	0.2898	0.2347	0.5874
ICTCSXRUN01	0.1648	0.2338	0.2497	0.2857	0.2347	0.2143	0.4245
UMDemailTLNR	0.1410	0.2015	0.1997	0.3388	0.2980	0.2357	0.5561
uwXSHUBS	0.1389	0.2028	0.1938	0.3551	0.2878	0.2449	0.5185
uogX06csnQE	0.1387	0.2046	0.2180	0.3061	0.2551	0.2071	0.5430
PITTPHFREQ	0.1117	0.1843	0.1744	0.3143	0.2857	0.2031	0.5085
allbasic	0.0996	0.1479	0.1409	0.3020	0.2429	0.1786	0.5233
sophiarun1	0.0934	0.1415	0.1322	0.3184	0.2449	0.1582	0.4646
SPlog	0.0781	0.1179	0.1470	0.2000	0.1694	0.1347	0.4265
l3s2	0.0714	0.0827	0.0820	0.3429	0.1755	0.0878	0.5840
body	0.0484	0.0809	0.1004	0.1224	0.1122	0.0918	0.2606

Table 5: Expert ranking with retrieval of a correct supporting document required. Runs in italics are manual runs.

Number of pooled		Experts	Tau correlation	
Experts	Documents	per Topic	No support	Support required
20	10	30.1	0.96	0.99
20	5	27.9	0.96	0.98
20	1	21.6	0.93	0.94
10	10	22.5	0.95	0.97
10	5	20.5	0.94	0.96
10	1	16.0	0.90	0.92

Table 6: Comparison of system rankings using pools of 20 or 10 experts and 10, 5, or 1 support documents per pooled expert, using the tau correlation to the official ranking.

3.3 Reducing pool size

As indicated above, the inclusion of support documents for experts caused the pools to be very large. Using these pools, we can examine if equivalent evaluation results could be obtained with smaller pools. The judged pools included the top 20 retrieved experts and the top 10 retrieved support documents for each candidate expert. In the process of this experiment, we discovered a bug in the support document pooling. The outcome of this bug was that if an expert was in the pool, the top 20 support documents were pooled even from a run that did not retrieve that expert in its top 10. This increased the size of the pools by a factor of 1.5 on average, and it seems likely that most of those documents were not relevant, simply because they came from less effective runs. After correcting this error and re-creating the relevance judgments based only on what should have been pooled, we found that nearly all the relevant experts found in the official pools were still present in the corrected version. The tau correlation to the official results was 1.0 for unsupported MAP and 0.99 for supported MAP. Thus, we have not changed the official reported results based on the original relevance judgments.

Starting from these corrected pools, we further reduced them by taking only the top 1, 5, or 10 supporting documents, and similarly by taking the top 10 experts only and the corresponding top 1, 5, or 10 supporting documents. Since the expert judgments were presumably informed by the supporting documents, we could not just apply the original expert judgment in the reduced pools. Instead, we used the following heuristic: if a document supporting expertise was retained in the reduced pool, we judged the candidate an expert. Similarly, if the reduced pool contained a document judged as indicating that the candidate was not an expert, we judged the candidate to not be an expert. If both supporting and detracting documents were in the reduced pool, we retained the original assessor's judgment of expertise. If no supporting or detracting documents were in the reduced pool, the candidate's expertise was labeled unknown.

Table 6 compares the system rankings based on the relevance judgments from these reduced pools to the official rankings reported above. For both supported and unsupported MAP, all reduced relevance judgment sets provide system rankings that are nearly identical to the official ranking.

An important concern when using small pools is that runs that did not contribute to their creation may be unfairly scored, because these runs are more likely to have retrieved candidates and support documents that are not in the pool. To gauge this effect, within each set of reduced pools we held out each group's runs in turn and measured them using the relevance judgments that would have been produced if their group had not contributed. Again, this works as in the reduced pools themselves; candidates that are only found by the held-out group are left unjudged, and holding out a group's unique support documents can change a judgment of candidate expertise.

We consider both changes in score, as well as how a group would have been ranked differently

Number of pooled		No support			Support required		
Experts	Documents	max	min	rank	max	min	rank
20	10	+0.0048	-0.0639	+0/-15	+0.0044	-0.0641	+5/-15
20	5	+0.0026	-0.0061	+0/-16	+0.0046	-0.0751	+5/-15
20	1	+0.0043	-0.1269	+0/-22	+0.0069	-0.1084	+6/-26
10	10	+0.0138	-0.0443	+6/-12	+0.0156	-0.0550	+8/-14
10	5	+0.0173	-0.0581	+5/-15	+0.0123	-0.0579	+7/-14
10	1	+0.0394	-0.1116	+5/-21	+0.0258	-0.1060	+11/-26

Table 7: Changes in MAP score and rankings when groups' runs are left out of the pools. "max" and "min" are the maximum and minimum MAP score difference. "rank" gives the largest movements up and down in the ranking when a group's runs are held out.

had it not contributed to the pool. Table 7 shows these results. "max" and "min" show the maximum and minimum change in MAP score among held-out groups. "rank" shows the largest moves up and down in the ranking. For example, when the pool is reduced to contain only 10 candidates per run and a single support document per candidate (10-1), one run drops 26 places in the supported experts ranking of 91 runs when all runs from that group are held out of that reduced pool. Note that this large change indicates that most runs scored very closely together; a change of -0.1060 in MAP covers more than a quarter of the ranking.

These results seem to indicate that the pools can be significantly reduced and still adequately measure the pooled runs, but that some caution should be exercised to ensure that the judgments are reusable by groups that did not participate. Reducing down to a single support document has a very large effect, greater than pooling fewer experts. Ten candidates with five support documents each is probably reasonable.

4 Conclusion

The second year of the enterprise track was very successful. We built a second set of topics for searching for discussions in mailing lists. We have also built a test collection for expert search. Taken together, the enterprise track collections are the first of their kind. While we still need to study their stability and reusability, we hope they will be a valuable resource for researchers.

An important lesson we have learned is that it can be difficult to situate information needs within an organization when you are not actually part of that organization. The topics largely give the impression of someone on the outside looking in, perhaps representative more of a new member of an organization rather than a veteran. When we began the track, we were concerned that the technical nature of the organization would be the chief obstacle to topic development. Now with topics created both by TREC participants and by NIST assessors, we appreciate that the greater challenge is to think of the information needs that people inside the organization have.

To that end, the collection will change in TREC 2007. The collection will be a snapshot of CSIRO, the Australian Commonwealth Scientific and Industrial Research Oganization. More importantly, the topics will be developed by employees at CSIRO. This will result in a topic set that reflects the range of information needs found within the organization.

5 Approaches

The following are descriptions of the approach taken by different groups. These paragraphs were contributed by participants and are intended to be a road map to their papers in the TREC proceedings. Below each group name is a list of their runs submitted to each task.

5.1 Beijing University of Posts and Telecommunications

Expert: PRISEXB, PRISEXR, PRISEXRM, PRISEXRMT

Candidates are ranked by their relevant description files. Each description file is constructed with the words co-occurred with a candidate, i.e., in the same window of text, in a document. Support documents are also ranked according their corresponding description files. Special data structures like headword and email are also considered to improve performance.

5.2 Case Western Reserve University

Expert: allbasic, basic, w1r1s1

This was Case Western Reserve University's first participation in TREC. We participated in the expert search task of the enterprise track. Our motivation for participation was our work developing an expert search capability for a prototype vertical digital library, MEMS World Online (`memsworldonline.case.edu`). For the expert search task we mostly relied on the email list portion of the W3C corpus. The emails are likely to be the most accurate indicator of an individuals expertise. Additionally, we give higher weight to response emails, which are also likely to be good indicators of expertise. We also used an additional weighting factor which is related the expertise of the individual's closely related colleagues in the social network extracted from the corpus. This is based on the intuition that the experts of the same topic are likely to work closely together. Finally, we used WordNet for synonyms in one run, though we did not expect much from this because of the technical nature of the task topics. We did not do any significant file preprocessing and only used automatic queries.

5.3 Chinese Academy of Sciences – ICT

Discussion: IIISRUN

Expert: ICTCSXRUN01 – 05

In this year, our team's research and experiments mainly focus on the mail list corpus and the link relationship amongst the candidates expert and other users. The W3C corpus includes a large archive of the W3C's mail lists. These lists are email forums for people who want to share information about W3C's research and projects. We can treat these forums as social networks. In our experiments, we find some interesting features of the community structures of these networks: In most of the mail lists, the candidate experts are not well connected. The social network in these mail lists can be divided into some communities which includes a few candidate experts and a lot of other users. The candidate experts are mostly in the center of their communities. And also, we use some link analysis approaches to rank the candidates in the social networks. In our experiments, we choose the PageRank algorithm and a revised HITS algorithm as link analysis methods. These approaches gives satisfying results in our experiments.

5.4 City University

Expert: ex3512, ex5512, ex5518, ex7512

A naive string matching algorithm is used to extract the full name and email addresses of identified experts, using a fixed window size (of 2000 characters), in order to build a profile for

those experts. We then index these profiles using Okapi, and used BM25 to rank the experts to generate our results.

5.5 DaLian University of Technology

Discussion: DUTDS1 – 4
Expert: DUTEX1 – 4

For email discussion search, we first preprocessed the cleaned W3C collections based on which an index was built by Indri (or Lemur). Then we handled the query topic in the same way of cleaning the documents, i.e. stripping the special character and stopping word. Ultimately, relevant documents were retrieved by Indri (or Lemur).

For expert search, we first created a correlative document pool for each candidate from the cleaned W3C collections and then gained the expert list and the support document with the pool. In the stage of correlative document pool generation, firstly, we collected the identities of each candidate, including his name, email, phone, nick, personal main page and so on. There were two stages in this process, automatic and manual. In the automatic we made several rules for identity extraction combining the technique of named identity recognition, then adjust and recruit the result in the manual stage.After candidate identity extraction was finished, an index was built based on the cleaned W3C collections and utilized the candidate identities to query. We singled out a number of words around the candidate identity to form the correlative document pool.

In the stage of expert list and supporting document generation, an index was built based on the correlative pool firstly. We attempt to compose the query in several ways for each topic and introduced the query to the Indri. The expert list was gained through the retrieved Indri score.Different from lasts year, every retrieved expert should be provided with his supporting documents which can explain why the candidate is an expert in this subject. Accordingly, we dealt with the correlative document pool. We took the (document ID-candidate ID) as the supporting document ID, in this way the correlative document pool of a candidate was divided into some supporting documents. Then we added the candidate identities to the original query and utilized Indri to gain the supporting documents of the expert.

5.6 L3S, University of Hannover

Expert: l3s1 – 4

We performed experiments on Expert Search in scope of Enterprise Track 2006. We based our technique solely on W3C mailing lists. The main assumption was that the author of an email is an expert on the subject addressed by the email. We tested 4 different heuristics with different threshold on the document score as well as the expert score. Using set of data-driven thresholds on similarity values, we cut off different number of experts per each query. One finding of our experiments was that complexity of the information need does not correlate with the number of relevant experts returned by the system. It was an interesting result, since normally the more specific your question, the less experts you expect. This result should be investigated more carefully, since definition of the task specificity is somewhat vague. It would be interesting to agree on one common scheme for task specificity definition in the expert search community.=20 We also scheduled more experiments with additional dataset, which we are creating in our group. This dataset will include real world documents, publications and wiki pages. The difference with W3C collections is that it could be enhanced with specific expert search interface and will allow tracking user logs while searching experts in it.

5.7 Lowlands Team

Expert: MAPCrelTret, MAPTrelCret, SP, SPlog

The lowlands team worked on the expert search task. We experimented with directly comparing two sets of document rankings: one for topics one for candidates. For each candidate we produce a ranked list of the 1000 most relevant documents based on a name+email address query. For each topic we produce a separate ranked list of the top 1500 most relevant documents. The intuition is that candidates for whom the document ranking has a high correlation with the ranking based on a given topic are likely to be experts for that topic. Experiments with various ways of producing the candidate based rankings and various ways of computing the correlation, showed that with a good document ranking for the candidates, good results can be obtained independent of the correlation method used.

5.8 Open University

Expert: kmiZHU1, kmiZhu2, kmiZhu4, kmiZhu5

Our group have used a two-stage language modeling approach consisting of a document relevance model and a window-based co-occurrence model in expert search. Document relevance measures the relevance of a document to a topic, and the co-occurrence model measures the relevance of an expert to a topic. Boolean query, span query, BM25, and TF/IDF are used for document relevance. There are mainly three innovative points in our group's approach. First, document authority in terms of their PageRanks is taken into account in the document relevance model, and the assumption is that more authoritative documents are linked or referenced more often by the others. Second, document internal structure is considered in the co-occurrence model. The occurrence of an expert's name in different parts of a document has influence on judging his/her relevance to a topic. We used templates of documents to segment these documents and consider structures of various documents, e.g., technical report, emails, and research papers. Third, we used incremental window sizes in the co-occurrence model. In selecting window sizes, small windows often lead to more accurate associations between experts but may miss some of them, while large windows often cover more associations to compensate small windows but may introduce noise. We gave higher weights to small window based than large window based relevance and aggregate their relevance together. Window sizes can reflect from phrase level, sentence level up to document level associations. In addition to the three points, partial match of queries, query construction from description and narrative of topics, and query construction by domain experts were also studied.

5.9 Queen Mary University of London

Expert: body, listbq, quotes, www

For Enterprise TREC, our group tried a strategy which integrates information retrieval with database management techniques. We use a probabilistic framework that allows us to evaluate expert finding strategies expressed in probabilistic variants of SQL and Datalog. Documents in the ETREC collection are parsed into a relational representation, to aid the integration of IR and DBMS. For the identification of experts, we assumed that some parts of emails in the collection are better at discriminating experts than others. We used different runs to check this claim, using only quotations, only bodies, or the whole email text for expert finding, and compared the performance of these different strategies.

5.10 Queensland University of Technology

Expert: qutbaseline, qutlmv2, qutmoreterms

We have participated in the expert search using the Terrier search engine for topic based retrieval, and then post-processed the top 100 documents to identify the experts. The concept of an expert was identified through the frequency with which the expert appears in the top 100 documents (emails, news, standards or drafts). The heuristic is pretty straight forward – one would expect a higher frequency for an expert in publication, citation, email discussion, etc. Furthermore, the persons appearing in the W3C standards or drafts as editors or authors should be experts. We did not have an opportunity to refine the selection to take account of indicative context. We based our expert selection on frequency alone without any attention to context or other details. The performance of the system was quite reasonable considering its simplicity. The system outperformed the median score when measured over all topics, but was not quite competitive enough relative to the best topic scores although it got close for several topics.

5.11 Ricoh Software Research Ctr.

Discussion: srcbds1 – 5
Expert: SRCBEX1 – 5

We participated in expert search and discussion search of Enterprise Track in TREC 2006. In the discussion search, we take advantage of the redundant pattern of emails to parse them according to their data structure. The collected pieces of information are subsequently stored in XML format and include the subject part, author part, sent time part, content part, quoted message part, greeting part and ad part. As the words in different parts are known to have different semantic weight, we use the so-called Field-Based weighting method to find relevant documents. We not only consider content relationships between the query and the target document but also non-content features such as time-line, mail thread, author, category and quoted chain. Tests showed that these non-content features are effective in improving the precision of discussion search. Our expert search consists of four features. Firstly, we make two kinds of data clean - webpage clean and candidate clean to adopt a profile-based document search. Core information is extracted from the W3C corpus such as the title, bolds, abstract, etc. Candidates are then matched with each web page and a profile is created for each candidate. Secondly, we use two variation weighting models, variation BM25 weighting model and DFR_BM25 weighting model. Query-based document length, not profile length, is used as document length in these weighting models to eliminate multiple topic noise. Query-based document length is the summation of lengths of extracted web pages that are relevant to the query. Thirdly, we use variation phrase weighting model to decease semantic confusion. Fourthly, field based two stage search method is used to make refined search. We demonstrate, on the basis of experiments, how these four approaches can effectively improve expert search.

5.12 Shanghai Jiao Tong University

Expert: SJTU1 – 4

In this research, we propose a new evidence-oriented framework to expert search. Here, the evidence is defined as a quadruple like (Query, Expert, Relation, Document). Each quadruple denotes that a "Query" and an "Expert", with a certain "Relation" between them, are found in a specific "Document". Within this framework, the task of Expert Search can be accomplished in three steps, namely, 1) evidence extraction: various kinds of co-occurrences between the expert and the query are extracted; 2) evidence quality evaluation: many novel factors like matching quality and context quality, are proposed as evidence quality evaluation; and 3) evidence merging: we proposed and compared two novel methods for evidence merging. The experimental results show that the new exploited evidences are quite useful and the evaluation of evidence quality improves the expert search significantly.The results also show that with cluster based merging, the result becomes even better.

5.13 Tsinghua University

Discussion: THUDSSUBPFSM, THUDSSUBPFSS, THUDSTHDM, THUDSTHDPFSM, THUDSTHDPFSS

Expert: THUPDDEML, THUPDDFBS, THUPDDL, THUPDDS, THUPDDSNEMS

Our expert finding system derives from that of last year, which first reorganize original documents to form PDDs, and then search and rank experts from these PDDs by employing retrieval model based BM2500 algorithm and bi-gram weighting. Our work this year focuses mainly on refinement of PDD documents and result reranking. We take advantage of email documents by producing Email-PDDs, appending Email subjects to original PDDs to form new PDDs, and combining search results of new PDDs and Email-PDDs. Regarding the result reranking stage, we have examined whether certain query-independent features – such as person activity and expert degree – help to find experts more accurately. Another new reranking approach we probed is to make use of social network, which is synthesized based on co-occurrences in web pages or email communications.

In Discussion Search task, several approaches have been probed. First, we discard useless and meaningless information in the email corpora to diminish the noise that affects the retrieval results. Then we examine the effectiveness of different field features in email such as quoted text and subjects of the email, some field features are emphasized by enforced as PFS (Primary Field Space) in our retrieval model. Finally we combined the adjacent serial emails to email threads and calculate the similarities of the single email and its threads respectively then integrate them together. Queries were constructed from the "query" field and "description" field. And all the experiments are base on our search engine TMiner.

5.14 University of Amsterdam

Discussion: UAmsBase, UAmsPOSBase, UAmsPOStQE, UAmsThreadQE

Expert: UvAbase, UvAPOS, UvAprofiling, UvAprofPOS

Following upon our last year's TREC Enterprise participation, we employ a standard language modeling setting for both tasks. Our aim for the discussion search task was to experiment with various query expansion techniques. Our first method employs blind relevance feedback, but instead of using the top ranked documents, we also include the contents of the accompanying threads. Our second method enriches the query by adding noun phrases from the description and narrative fields. We also experimented with combining the outcomes of the different approaches. Results indicate that adding terms from the description and narrative fields helps in most cases but not all. Thread-based query expansion did not deliver the desired results, due to topic drift. As to the expert search task, our baseline method calculates the probability of a candidate being an expert given the query topic. This probability is estimated by iterating over all documents that are associated with the given person. Moreover, we introduce the topical profile of an individual, which reflects the person's competency over a set of knowledge areas. The expert search topics were used as knowledge areas, and the topical profile of each W3C candidate was calculated. A rank-based combination of expert finding and profiling methods resulted in remarkable improvements over the baseline.

5.15 University of Glasgow

Expert: uogX06csnP, uogX06csnQE, uogX06csnQEF, uogX06ecm

In our participation in the Enterprise Track, we aim to develop our novel voting model for expert search. Our newly-proposed approach models expert search as a voting process. In our model, a candidate's expertise is represented by a profile, which is a set of documents associated to the candidate. Then, using the ranked list of retrieved documents for the expert search query,

we propose that the ranking of candidates can be modeled as a voting process, from the retrieved documents to the profiles of candidates. The votes for each candidate are then appropriately aggregated to form a ranking of candidates, taking into account the number of voting documents for that candidate, and the topicality of the voting documents. Our voting model is extensible and general, and is not collection or topics dependent.

This year in TREC, we test two new approaches for appropriately aggregating the votes for candidates. Moreover, we integrate a new component into the model that takes into account the candidate's profile length. Finally, we test a selection of approaches to increase the accuracy of the voting documents.

5.16 University of Illinois at Urbana-Champaign

Expert: UIUCe1, UIUCe2, UIUCeFB1, UIUCeFB2

We submitted four automatic runs, all using the title field of a topic and the whole corpus. Our goal is test the effectiveness of a new language model for expert retrieval. The new language model is based on the model 2 proposed in (Balog et al., 2006) with the following three extensions: (1) We model the document-candidate association using a mixture model that allows for putting different weights on matching the email and matching the name of a candidate. Thus we have a complete unigram language model for this task. (2) We use the count of email matches in the supporting documents for a candidate to define a prior on candidates such that a candidate whose supporting documents have many email matches would be favored. (3) We perform topic expansion and generalize the language model from computing the likelihood to computing the KL-divergence.

5.17 University of Maryland

Expert: UMDemailTLNR, UMDemailTTL, UMDthrdTTLDS, UMDthrdTTL, UMDthrdTTLNR

We have adopted a simple unsupervised approach that focuses only on mailing lists as the source of evidence of candidate expertise. The system first retrieves a set of emails or threads that are relevant to the topic and scores the candidates based on references in the headers and mentions in the text to their names and email addresses in the retrieved set. The credit given by each reference or mention is weighted according to (1) the retrieval similarity (to the topic) score of the email where the reference appears, and (2) in which field (headers, new text, quoted, etc.) in that email it appears.

5.18 University of Massachusetts

Discussion: UMaTDMixThr, UMaTiMixHdr, UMaTiMixThr, UMaTiSmoThr

Expert: UMaTDFb, UMaTiDm, UMaTNDm, UMaTNFb

This year the University of Massachusetts took part in both tasks of the Enterprise track. For the DS task we compare two methods for incorporating thread evidence into the language models of email messages. To group emails by thread we used the *all-in-reply-to* list provided by William Webber, concatenating the text of related messages.

One approach for incorporating thread context is to estimate a language model of the thread and interpolate it with the smoothed language models of other email components (header and mainbody). We use Dirichlet smoothing and automatically set the α parameter equal to the average component length. An alternative way to take advantage of thread information is to use it as a background model for smoothing email components. The idea is that threads would provide a more reasonable fallback distribution than a word distribution for general English. Our experimental results show that smoothing with a thread-based fallback model is more effective

than smoothing with a general collection model. However, constructing a mixture of language models from header, main body and thread text is more effective.

Our approach to the ES task represents candidate experts as mixtures of language models from associated documents and then ranks candidates according to query likelihood. Since the candidate representations are probability distributions over terms, we can build richer models by interpolating models estimated from different subcollections or different types of documents, or different entity definitions; in short, retrieval settings representing different descriptions (aspects) of a person entity. For example, we use two subcollections (www and lists), and two definitions (full name and last name). This model also preserves the information inherent in individual documents, such as structure and term proximity. Therefore we can use document retrieval techniques to capture higher-level language features. We use pseudo-relevance feedback and phrase expansion.

5.19 University of Ulster and St. Petersburg State University

Expert: sophiarun1 – 3

The SOPHIA group used the Contextual Document Clustering algorithm to cluster the W3C document corpus (documents from www and lists catalogs) into hundreds of thematically homogeneous clusters. Given a topic, the most relevant clusters were used to select experts for that topic. The expert relevancy score was calculated based on the number of mails sent by the expert from within the relevant clusters and similarities between these mails and the topic.

5.20 University of Waterloo

Discussion: uwTbaseline, uwTDbaseline, uwTDsubj, uwTsubj

Expert: uwXSHUBS, uwXSOUT, uwXSPMI

For the discussion search task, we hypothesized that the author's of an email tend to give their subjective opinion about the topic in discussion. In this year's discussion search track, we tested this hypothesis by re-ranking the email lists based on the presence of certain subjective adjectives in the proximity of the query words.

Experts, people who are knowledgeable about a given topic, tend to associate themselves with the topic over certain period. For expert search, in one approach, we estimated the association with the topic by studying the patterns in the mailing lists. We used graph-based ranking algorithm like HITS algorithm and PageRank to rank the candidates. In other approach, we estimated the expertise using statistical measures like mutual information etc, b/n the candidate and the topic.

References

Krisztian Balog, Leif Azzopardi, and Maarten de Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–50, Seattle, WA, August 2006.

Nick Craswell, Arjen P. de Vries, and Ian Soboroff. Overview of the TREC 2005 enterprise track. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, November 2005. URL `http://trec.nist.gov/pubs/trec14/t14_proceedings.html`.

Yejun Wu, Douglas Oard, and Ian Soboroff. An exploratory study of the w3c mailing list test collection for retrieval of emails with pro/con arguments. In *Proceedings of the Third*

Conference on Email and Anti-Spam, Mountain View, CA, July 2006. URL `http://www.ceas.cc/2006/26.pdf`.

Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pages 102–111, Arlington, Virginia, November 2006.

TREC 2006 Genomics Track Overview

William Hersh[1], Aaron M. Cohen[1], Phoebe Roberts[2], Hari Krishna Rekapalli[1]

[1]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA, [hersh, cohenaa, rekapall]@ohsu.edu
[2]Biogen Idec Corp., Boston, MA, USA, Phoebe.Roberts@biogenidec.com

The TREC Genomics Track implemented a new task in 2006 that focused on passage retrieval for question answering using full-text documents from the biomedical literature. A test collection of 162,259 full-text documents and 28 topics expressed as questions was assembled. Systems were required to return passages that contained answers to the questions. Expert judges determined the relevance of passages and grouped them into aspects identified by one or more Medical Subject Headings (MeSH) terms. Document relevance was defined by the presence of one or more relevant aspects. The performance of submitted runs was scored using mean average precision (MAP) at the passage, aspect, and document level. In general, passage MAP was low, while aspect and document MAP were somewhat higher.

1. Introduction

The goal of most information retrieval (IR) systems is to retrieve documents that a user might find relevant to his or her information need. In contrast, the goal of most information extraction (IE) or text mining (TM) systems is to process document text to provide the user with one or more "answers" to a question or information need (Cohen and Hersh, 2005; Roberts, 2006). We propose that what many information seekers, especially users of the biomedical literature, really desire is something in the middle, i.e., a system that attempts to provide short, specific answers to questions and put them in context by providing supporting information and linking to original sources (Hersh, 2005). This motivated us to go beyond the ad hoc retrieval task from previous years of the TREC Genomics Track (Hersh, Cohen et al., 2005; Hersh, Bhupatiraju et al., 2006).

For the TREC 2006 Genomics Track, we developed a new task that focused on retrieval of short passages (from phrase to sentence to paragraph in length) that specifically addressed an information need, along with linkage to the location in the original source document. Topics were expressed as questions and systems were measured on how well they retrieved relevant information at the passage, aspect, and document levels. Systems were required to return passages linked to source documents, while relevance judges not only rated the passages, but also grouped them by aspect. For this task, aspect was defined similar to its definition in the TREC Interactive Track aspectual recall task (Hersh, 2001), representing answers that covered a similar portion of a full answer to the topic question. We also drew upon experience in passage retrieval from the previous TREC High Accuracy Retrieval from Documents (HARD) Track (Allan, 2003; Allan, 2004).

2. Document collection

The documents for this year's task came from a new full-text biomedical corpus. We obtained permission from a number of publishers who use Highwire Press (www.highwire.org) for electronic distribution of their journals. They agreed to allow us to include their full text in HTML format, which preserved formatting, structure, table and figure legends, etc. The document collection was derived from 49 journals and were obtained by a Web crawl of the Highwire site, with post-processing to eliminate as much non-article material as we could. The full collection contained 162,259 documents. The collection was about 12.3 GB when uncompressed. Appendix 1 lists the journals and number of documents from each.

Several notable issues were uncovered when the collection was compiled:

- The collection was not complete from the standpoint of each entire journal. That is, there were some articles that appeared in the journal but did not make it into our collection. This was acceptable to us, since we viewed the collection as a closed and fixed collection.
- Some of the PMIDs were incorrect, emanating from errors in the URLs linking to Pubmed in the source data from Highwire Press.
- Some of the HTML files were empty or nearly empty (i.e., only contained a small amount of meaningless text). Some of this was due to errors in our processing, but most was related to the incorrect or ambiguous links on the Highwire site and in the HTML documents themselves. We kept these files in the collection since they were small and unlikely to have any relevant passages.

We also created a text file, metadata.txt (Windows ASCII format, 11.9 MB), which listed the original URL of the article, the file name in our collection, and the file size in kilobytes. The name of each document file was its PMID plus the extension ".html", which facilitated accessing the associated MEDLINE record.

In addition to the full-text data, the National Library of Medicine (NLM) provided us with both ASCII and XML formatted collections of all the MEDLINE records for the full-text documents in our Highwire collection. We identified 1,767 instances (about 1% of the 162K documents) where the Highwire file PMID was invalid. We investigated the problem and found that for all of instances we checked, the problem was in the original Highwire HTML file having an incorrect PMID in the link to the PubMed record. In other words, the error was inherent in the Highwire data, and not introduced as a result of our processing.

Another file made available to participants was legalspans.txt. This file contained all "legal spans" for all documents in the collection. Legal spans were defined as any contiguous text >0 characters in length not including any HTML paragraph tags, defined as any tag that started with <P or </P (case insensitive). There were a total of 12,641,127 legal spans in the collection. We used these spans to define allowed passages in the pooling and evaluation process, and to limit the size of the passages that needed reviewing by the expert judges

Retrieved passages could contain any span of text that did not include any part of an HTML paragraph tag (i.e., one starting with <P or </P). Because there was some confusion about the different types of passages, we defined the following terms:

- Nominated passage - This was the passage that systems nominated in their runs and were scored in the passage retrieval evaluation. To be legal, these passages had to be a subset of a maximum-length legal span.
- Maximum-length legal span - These were all the passages obtained by delimited the text of each document by the HTML paragraph tags. As noted below, nominated passages could not cross an HTML paragraph boundary. So these spans represented the longest possible passage that could be designated as relevant. As also noted below, we built pools of these spans for the relevance judges. The judges were given the plain text from the entire maximum-length legal span, even if no system nominated the entire span. However, the judges did not need to designate the entire span as relevant, and were able to select just a part of the span as the relevant passage. Each maximum length span was identifier by a triple value of (PMID, offset, length).
- Relevant passage - These were the spans that the judges designated as definitely or possibly relevant. These were portions of the original HTML files, represented by the value triple: PMID, offset, and length. These spans may or may not include HTML markup tags, depending upon whether these tags were inside the relevant answer passages designated by the experts.

The following should also be noted about the maximum-length legal spans:

- The first and last spans were delimited at the beginning and end of the file respectively.
- Other HTML tags (e.g., <B>) could occur within the spans.
- "Empty" (zero character) spans were not included.

3. Topics

The topics for the 2006 track were expressed as questions. They were derived from the set of biologically relevant questions based on the Generic Topic Types (GTTs) developed last year for the 2005 track. These questions each had one or more aspects that were contained in the literature corpus (i.e., one or more answers to each question). A few things should be noted about the topics for 2006:

- Even though the questions were derived from the 2005 track topics, many of them changed, some substantially.
- Groups were instructed that if their systems made use of knowledge about the 2005 topics, then they needed to classify their 2006 runs as interactive, even if they only used automated methods in 2006.
- The official topics were the text of the questions in the text file that was provided. We also provided an Excel spreadsheet, and corresponding PDF, which showed the 2005 topics from which the 2006 topics were derived. However, the information from the 2005 questions was for reference only, and was not to be considered part of the 2006 data.

The questions (and GTTs) all had the general format of containing one or more biological objects and processes and some explicit relationship between them:

Biological object (1..many) ← relationship → Biological process (1..many)

The biological objects might be genes, proteins, gene mutations, etc. The biological process could be physiological processes or diseases. The relationships could be anything, but were typically verbs such as *causes*, *contributes to*, *affects*, *associated with*, or *regulates*. We determined that four out of the five GTTs from 2005 could be reformulated into the above structure, with the exception of the first GTT that asked about procedures or methods. The patterns for doing this from the GTTs were based on the examples in Table 1. The topics for the 2006 track are listed in Table 2.

4. Submissions

Submitted runs could contain up to 1000 passages per topic that were predicted to be relevant to answering the topic question. Passages had to be identified by the PMID, the start offset into the text file in characters, and the length of the passage in characters. The first character of each file was defined to be at offset zero.

Passages were required to be contiguous and not longer than one paragraph. As described above, this was operationalized by prohibiting any passage from containing HTML markup tags, i.e., those starting with <P or </P. Any passages containing these tags were ignored in the judgment pooling process but not omitted from the scoring process. (In other words, not counted as potentially relevant for pooling but counted as retrieved for scoring.) Each participating group was allowed to submit up to three official runs, all of which were used for building pools. Each passage was required to be assigned a corresponding rank number and value. The rank number, starting at one and ascending, was used to order nominated passages for rank-based performance computations.

Table 1 - Generic topic types used in the TREC 2006 Genomics Track.

GTT	Question Pattern	Example
Find articles describing the role of a gene involved in a given disease.	What is the role of gene in disease?	What is the role of DRD4 in alcoholism?
Find articles describing the role of a gene in a specific biological process.	What effect does gene have on biological process?	What effect does the insulin receptor gene have on tumorigenesis?
Find articles describing interactions (e.g., promote, suppress, inhibit, etc.) between two or more genes in the function of an organ or in a disease.	How do genes interact in organ function?	How do HMG and HMGB1 interact in hepatitis?
Find articles describing one or more mutations of a given gene and its biological impact.	How does a mutation in gene influence biological process?	How does a mutation in Ret influence thyroid function?

Table 2 - Topics for TREC 2006 Genomics Track.

<160>What is the role of PrnP in mad cow disease?
<161>What is the role of IDE in Alzheimer's disease
<162>What is the role of MMS2 in cancer?
<163>What is the role of APC (adenomatous polyposis coli) in colon cancer?
<164>What is the role of Nurr-77 in Parkinson's disease?
<165>How do Cathepsin D (CTSD) and apolipoprotein E (ApoE) interactions contribute to Alzheimer's disease?
<166>What is the role of Transforming growth factor-beta1 (TGF-beta1) in cerebral amyloid angiopathy (CAA)?
<167>How does nucleoside diphosphate kinase (NM23) contribute to tumor progression?
<168>How does BARD1 regulate BRCA1 activity?
<169>How does APC (adenomatous polyposis coli) protein affect actin assembly
<170>How does COP2 contribute to CFTR export from the endoplasmic reticulum?
<171>How does Nurr-77 delete T cells before they migrate to the spleen or lymph nodes and how does this impact autoimmunity?
<172>How does p53 affect apoptosis?
<173>How do alpha7 nicotinic receptor subunits affect ethanol metabolism?
<174>How does BRCA1 ubiquitinating activity contribute to cancer?
<175>How does L2 interact with L1 to form HPV11 viral capsids?
<176>How does Sec61-mediated CFTR degradation contribute to cystic fibrosis?
<177>How do Bop-Pes interactions affect cell growth?
<178>How do interactions between insulin-like GFs and the insulin receptor affect skin biology?
<179>How do interactions between HNF4 and COUP-TF1 suppress liver function?
<180>How do Ret-GDNF interactions affect liver development?
<181>How do mutations in the Huntingtin gene affect Huntington's disease?
<182>How do mutations in Sonic Hedgehog genes affect developmental disorders?
<183>How do mutations in the NM23 gene affect tracheal development?
<184>How do mutations in the Pes gene affect cell growth?
<185>How do mutations in the hypocretin receptor 2 gene affect narcolepsy?
<186>How do mutations in the Presenilin-1 gene affect Alzheimer's disease?
<187>How do mutations in familial hemiplegic migraine type 1 (FHM1) gene affect calcium ion influx in hippocampal neurons?

Each submitted run was submitted in a separate file, with each line defining one nominated passage using the following format based loosely on trec_eval. Each line in the file had to contain the following data elements, separated by white space:

- Topic ID - from 160 to 187.
- Doc ID - name of the HTML file minus the .html extension. This was the PMID that was designated by Highwire, even though this may not have been the true PMID assigned by the NLM (i.e., used in MEDLINE). But this was the official identifier for the document within the corpus.
- Rank number - rank of the passage for the topic, starting with 1 for the top-ranked passage and preceding down to as high as 1000.
- Rank value - system-assigned score for the rank of the passage, an internal number that should descend in value from passages ranked higher.
- Passage start - the character offset in the Doc ID file where the passage begins, where the first character of the file is offset 0.
- Passage length - the length of the passage in 8-bit ASCII characters.
- Run tag - a tag assigned by the submitting group that should be distinct from all the group's other runs.

Because of the complex nature of this year's task, and most groups' not having a system in place before release of the topics, the classification of runs was complicated. "Usual" TREC rules (detailed at http://trec.nist.gov/act_part/guidelines/trec8_guides.html) would ordinarily categorize runs as follows:

- Automatic - no human modification of topics into queries for a system whatsoever.
- Manual - human modification of queries entered into a system but no modification based on results obtained (i.e., you cannot look at the output from your runs to modify the queries).
- Interactive - human interaction with the system, including modification of the queries after viewing the output (i.e., you look at the output from the topics and corpus and adjust your system to produce different output).

However, because we reused topics (with modification, sometimes substantial) from 2005, and because people were building systems up to the last minute, we adopted the following rules to be applied to classification of runs:

- If a group made any tuning or optimization of their retrieval system based on last year's topics, then their run should have been categorized as interactive this year, even if they did everything else in an automated fashion.
- If a group made any human generated modifications to the topic statements or their system for queries entered into their system, then the run should have been classified as manual.
- If groups made any modifications to the topic statements or their system for the queries entered into it based on looking at the output of passages and/or documents, then their run should have been classified as interactive.

As with many TREC tasks, groups were allowed to manually modify topics to create their queries to their systems. In addition, they were allowed to consult outside resources on the Web (e.g., gene databases), but only in a fully automated fashion. In other words, the original queries could be manually modified, but interaction with external resources could only be done in an automated fashion. For example, if a system pulled information from SOURCE, GenBank, or any other resource, the query to those sources and the information obtained from them had to be done in an automated way, i.e., without manual intervention.

5. Relevance assessments

a. Pooling

There were 92 submitted runs, with each nominating up to 1000 passages over 28 topics. Given our resources, this was far too much data to perform an exhaustive expert evaluation. Instead, we used a pooling method, similar to

that used by other document retrieval tasks in TREC.

For each topic a separate pool of passages was created for expert judging. Each ranked and submitted passage consisted of a (PMID, offset, length) triple, which was mapped to its corresponding maximum-length span, also identified as a (PMID, offset, length) triple. These spans were distributed in the legalspans.txt made available before submissions were due. Then, for each topic, pooling was done by taking the top ranked maximal legal span from each submitted run in a round-robin fashion, until the topic pool contained 1000 unique spans. In other words, the top ranked passage was taken from each submitted run, and then the second ranked passage if not yet included, and so on, until the submissions were exhausted or a pool contained 1000 spans.

To consistently subdivide source documents into shorter passages, the HTML <P> tag was used to approximate splitting up the document into paragraphs; as noted above we called these maximum-length legal spans. Likewise, legal submitted passages were limited to not include any HTML <P> tags. By definition, maximal legal spans did not overlap. Therefore, any legally submitted passage would have to be either a maximal legal span or a subset of exactly one maximal legal span.

In addition to HTML <P> tags, additional markup characters were embedded in the text, hampering the readability (thought they generally rendered well in a browser). Maximal legal spans generated in the previous step were converted to plain text by removing the HTML markup. This allowed the judges to concentrate on the content of the passages instead of having to deal with erratic formatting issues. Despite the attempt to remove HTML formatting, plain text was not fully restored to publication quality. Common modifications included loss of inline images that represented characters such as Greek symbols, and lack of conversion of HTML entity codes to more easily readable plain text punctuation characters such as ampersands At times, for some judges, these changes proved to be a distraction.

The plain text content from the pooled spans was then imported into an Excel spreadsheet. Columns were added to allow easy relevance judging. A drop down menu was provided to set the relevance of each passage, and cells were provided for the judges to cut-and-paste relevant plain text from the maximal legal span text field into an "answer text" field. Another column was provided for judges to cut-and-paste MeSH terms corresponding to relevant passage aspects. To make the Excel forms more user friendly for the judges, hyperlinks were added to the PubMed record for the PMID for the journal article for each passage, and also to enable quick access to the PubMed MeSH browser.

b. Judging

Relevance judges were provided with guidelines and a one-hour training course to improve the judging process. As this year's track was developed by the steering committee, the question and answer nature of the task raised discussion about what constituted a complete answer, prompting development of guidelines for dealing with anaphora and abbreviations to benefit participants and judges alike. In addition, the guidelines offered a brief introduction to MeSH, and tips for taking advantage of Excel features to monitor consistency and completion. Nine judges participated, and they were provided with an email list to discuss issues as they came up.

To assess relevance, judges were instructed to break down the question into required elements (e.g., the biological entities and processes that make up the GTT) and isolate the minimum contiguous substring that answered the question. In general, a passage was definitely relevant if it contained all required elements of the question and it answered the question. A passage was possibly relevant if it contained the majority of required elements, missing elements were within the realm of possibility (i.e. more general terms are mentioned that probably include the missing elements), and it possibly answered the question.

It was possible for a judge to designate any number of relevant passages from an individual article. It was also possible for a judge to designate multiple non-overlapping relevant passages from an individual pooled span. The judges evaluated the text of the maximum-length legal span for relevance, and identified the portion of this text that contained an answer, hereafter called the gold standard passage. This could be all of the text of the maximum legal span, or any contiguous substring. It was possible that one maximum legal span could contain two or more separate gold standard passages. Judges were instructed to duplicate rows with more than one gold standard passage, and process each row independently. Judges were not shown how many systems had retrieved each maximum-length span. Appendix 2 shows the number of maximum-length legal spans where part or all of the span was judged as definitely or possibly relevant; the remainder were counted as not relevant.

Relevance judges next determined the "best" answer passages and grouped them into related concepts. The judges then assigned one or more Medical Subject Headings (MeSH) terms (possibly with subheadings) to capture similarities and differences among retrieved passage aspects. We originally considered using Gene Ontology (GO) terms for this purpose, but an early analysis by our genomics domain expert determined that GO lacked sufficient coverage in many areas needed for this task and MeSH terms alone would provide sufficient coverage.

Judges assigned MeSH term-based aspects to each gold standard passage. They were instructed to use the most specific MeSH term, with the option of adding subheadings, similar to the NLM literature indexing process. If one term was insufficient to denote all aspects of the gold standard passage, judges assigned additional MeSH terms. All passages judged as definitely or possibly relevant were required to have a gold standard passage and at least one MeSH term.

A total of six topics were selected randomly for judgment in duplicate: 160, 165, 176, 179, 181, and 185. (We hoped to have more topics judged in duplicate but were unable to recruit judges for additional work.) Table 3 shows the agreement of the original and duplicate judges, where agreement indicates that any part of a maximum-length legal span was judged as relevant or not. The kappa statistic was calculated to assess chance-corrected inter-rater agreement. For five of the topics, the kappa statistic indicated "good" inter-rater agreement, with a value of 0.60. For topic 181, however, the kappa statistic was poor, with a value of 0.028. This outlier brought the overall kappa value for the six topics down to 0.032. What happened for topic 181 was that one judge interpreted relevance to the question very broadly and the other very narrowly. Table 4 shows the agreement of original and duplicate judges for MeSH terms assigned for aspects, which shows an even poorer rate of agreement. (Kappa could not be calculated due to the inability to calculate the number of MeSH terms not assigned.)

c. Processing

The final result of the judging process was a set of filled-out forms in Excel spreadsheet format. Each spreadsheet corresponded to the judged passages for one topic, one row per passage. If a passage was marked "Not" relevant, no further processing needed to be done, as this passage was not included in the gold standard. Passages marked "Definitely" or "Possibly" relevant were treated as relevant for purposes of the gold standard. The "Definitely" and "Possibly" relevant passages also had two additional associated data items: the relevant answer text cut and pasted from the maximal legal span, and a list of pipe character-separated MeSH terms.

The text and MeSH data associated with the relevant passages was processed to create a set of gold standard passages for each topic. Each gold standard passage consisted of the PMID of the document that the passage was from, the starting character offset, the length of the gold standard passage, and a list of pipe character-separated MeSH terms corresponding to the aspects for that passage.

Table 3 - Agreement of original and duplicate judges for relevant passages, where agreement indicates that any part of a maximum-length legal span was judged as relevant or not.

	Five topics (not including 181)		**Six topics (including 181)**	
	Duplicate judge relevant	Duplicate judge not relevant	Duplicate judge relevant	Duplicate judge not relevant
Original judge relevant	234	228	253	789
Original judge not relevant	53	4485	53	4905

Table 4 - Agreement of original and duplicate judges for MeSH terms assigned. (The cell where neither assigned in undefined.)

	Five topics (not including 181)		**Six topics (including 181)**	
	Duplicate judge assigned	Duplicate judge did not assign	Duplicate judge assigned	Duplicate judge did not assign
Original judge assigned	83	730	90	2407
Original judge did not assign	632	N/A	652	N/A

The starting character offset and length of the gold standard passage in the HTML journal article file was determined by an automated process. Using a dynamic programming algorithm similar to the third stage alignment step in BLAST (McGinnis and Madden, 2004), the relevant answer text selected by the expert judge was aligned with the text of the corresponding maximum length span in the HTML file in order to determine the best overlap. This step had the effect of finding the plain answer text in the HTML file, accounting and skipping over any intervening HTML markup. The starting offset into the HTML file, along with the length in characters in the HTML file matching up with the answer text was taken to be the gold standard passage for that judgment.

As noted above, judges assigned MeSH terms to designate the aspects of a complete topic answer that were addressed by each relevant gold standard passage. This allowed grouping of answer passages and estimatation of the performance of systems in providing a complete answer. Ideally, the MeSH terms provided by the expert judges would have been copied from the MeSH browser without error. However, an additional processing step was necessary because of several types of variation. First, sometimes judges typed in MeSH terms instead of cut and pasting them. Spelling errors were introduced, and these needed to be corrected. A second type of error was created by judges using a MeSH entry term instead of the official MeSH descriptor. These entry terms needed to be mapped to the official term. A few errors were introduced by the judges when non-MeSH terms were used, these needed to be mapped to the closest official MeSH term.

Except for the spelling variations, judges were consistent within a topic, and so the MeSH term variations did not have any effect on the final results. However the MeSH assignments were normalized by mapping to the official MeSH descriptor in order to improve the overall quality and reusability of the test collection. A table driven program was created to fix these errors and map all aspects to official MeSH terms. The table was reviewed by our lead biological expert (P.R.) before finalizing the gold standard aspects. All MeSH terms were also normalized to upper case. Subheadings were preserved if used by the relevance judges.

After mapping the answer text to the HTML source documents and correcting variation in the

MeSH terms, the gold standard passages for each topic were combined into a single file. This file contained 3451 gold standard passages, giving the topic identifier, the source document PMID, the starting offset and length of the relevant passage, and a list of pipe character separated normalized MeSH terms.

Appendix 3 lists the number, average length, and standard deviation of passages per topic as well as the number of aspects per topic. Table 5 shows the minimum, mean, median, and maximum for the topics of these values. It is clear that there is significant variation among the topics as far as number of relevant passages in the literature corpus, the length of those passages, and the number of aspects per topic that were found by the judges. Note that two topics, 173 and 180, had no relevant passages.

6. Performance Measures

For this year's track, there were three levels of retrieval performance that we measured: passage retrieval, aspect retrieval, and document retrieval. Each of these provided insight into the overall performance for a user trying to answer the given topic questions. Each was measured by some variant of mean average precision (MAP). Because this was a new task, and uncharted research territory, we decided to measure the three types of performance separately. We did not propose any summary metric to grade overall performance, but instead wished to examine each aspect of performance in a way that was both as meaningful and straightforward as we could at our current level of experience with this task.

a. Passage-level MAP

This measure used a variation of MAP, computing individual precision scores for passages based on character-level precision, using a variant of a similar approach used for the TREC 2004 HARD Track (Allan, 2004). For each nominated passage, the number of characters that overlapped with those deemed relevant by the judges in the gold standard was determined. For each relevant retrieved passage, precision was computed as the fraction of characters overlapping with the gold standard passages divided by the total number of characters included in all nominated passages from this system for the topic up until that point. Similar to regular MAP, remaining relevant passages that were not retrieved (no overlap with any nominated passages) were added into the calculation as well, with precision set to 0 for these relevant non-retrieved gold standard passages. Then the mean of these average precisions over all topics was calculated to compute the MAP for passages. Note that this measure is essentially the fraction of retrieved characters that are part of an answer to the topic question.

b. Aspect-level MAP

Aspect retrieval was measured using the average precision for the aspects of a topic, averaged across all topics. To compute this, for each submitted run, the ranked passages were transformed to two types of values, either the aspect(s) of the gold standard passage that the submitted passage overlapped with or the value "not relevant". This resulted was a ranked list, for each run and each topic, of lists of aspects per passage, Non-relevant passages had empty lists of aspects. Because we were uncertain of the utility for a user of a repeated aspect (e.g., same aspect occurring again further down the list), we discarded these from the output to be analyzed. For the remaining aspects of a topic, we calculated MAP similar to how it is calculated for documents, with the additional wrinkle that a single passage may have associated with it multiple aspects. Therefore the precision for the retrieval of each aspect was computed as the fraction of relevant passages for the retrieved passages up to the current passage under consideration. These fractions at each point of first aspect retrieval were then averaged together to compute the average aspect precision. Relevant passages that did not contribute any new aspects to the aspects retrieved by higher ranked passages were removed from the ranking. Taking the mean over all topics produced the final aspect-based MAP.

Table 5 - Range and central tendency of relevant passages, their length, and distinct aspects per topic.

Measure	Relevant passages per topic	Mean relevant passage length	Distinct aspects per topic
Minimum	3	27	7
Mean	35	400	22
Median	133	229	30
Maximum	593	6928	96

c. Document-level MAP

For the purposes of this measure, any PMID that had a passage associated with a topic ID in the set of gold standard passages was considered a relevant document for that topic. All other documents were considered not relevant for that topic. System run outputs were collapsed by PMID document identifier, with the documents appearing in the same order as the first time the corresponding PMID appeared in the nominated passages for that topic. For a given system run, average precision was measured at each point of correct (relevant) recall for a topic. The MAP was the mean of the average precisions across topics.

Two topics, 173 and 180, had no relevant passages. These were not included in the scoring for any of the three measures.

7. Results

Information about each run is listed in Appendix 4, including a brief system description provided by the group. The results from all submissions are provided in Appendix 5. A summary of the medium and maximum run results by run type is shown in Figure 1. The best results per topic are seen in Figure 2. In general, document MAP scores are highest, followed by aspect, and then passage, although these scores are not directly comparable since they measure precision at recall of different things. There was a general, though far from perfect, correlation between passage, aspect, and document MAP, as shown in Figure 3. As seen in many TREC-style evaluations and demonstrated in Figure 4, statistical significance, based on pair-wise comparison with the top-ranking score in an ANOVA model, was not achieved for any measure until well down the ranked list of runs.

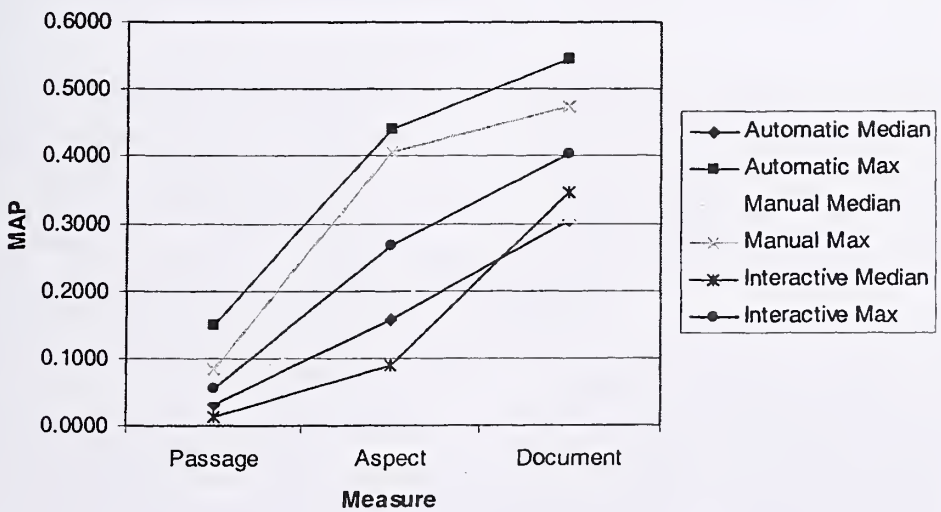


Figure 1 - MAP for all runs and those categorized as automatic, manual, and interactive.

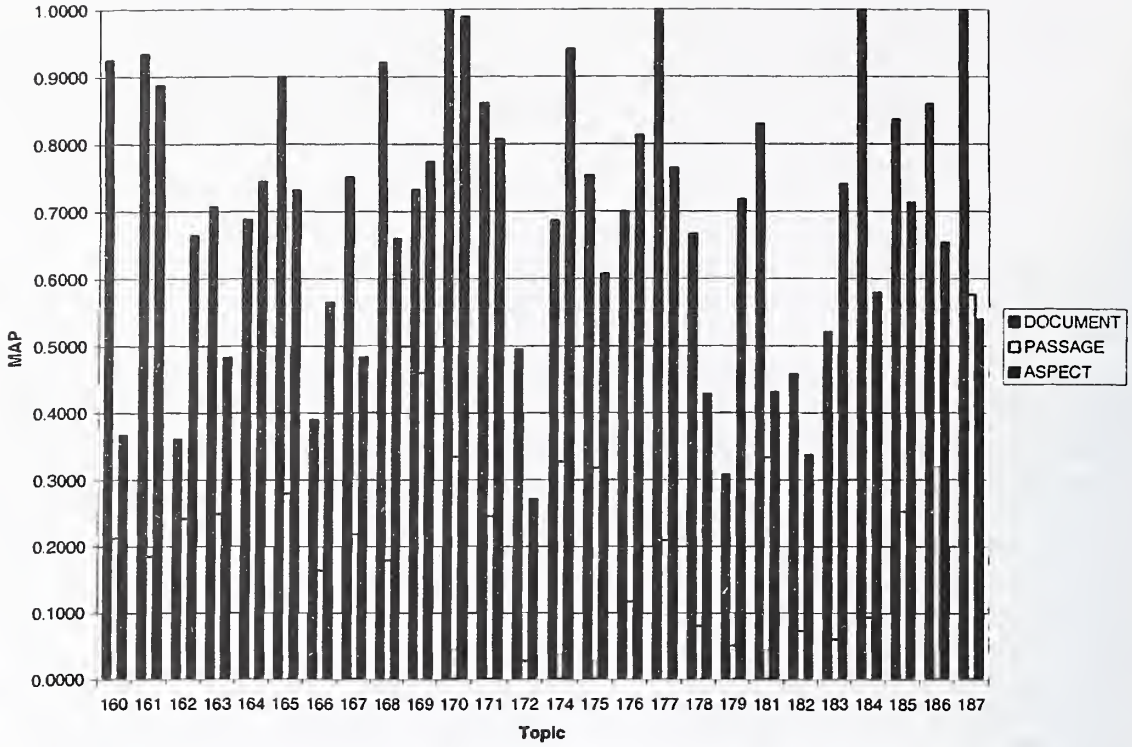


Figure 2 - Best results per topic

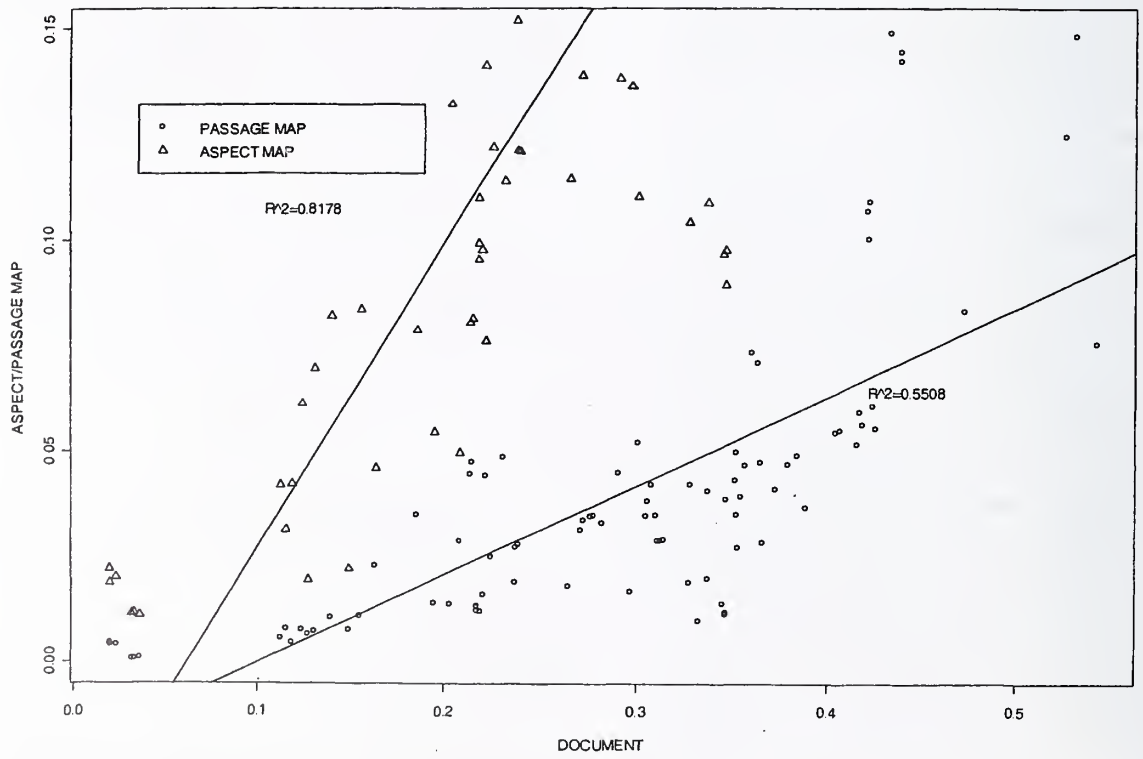


Figure 3 - Plot of passage and aspect MAP versus document MAP for all submitted entries.

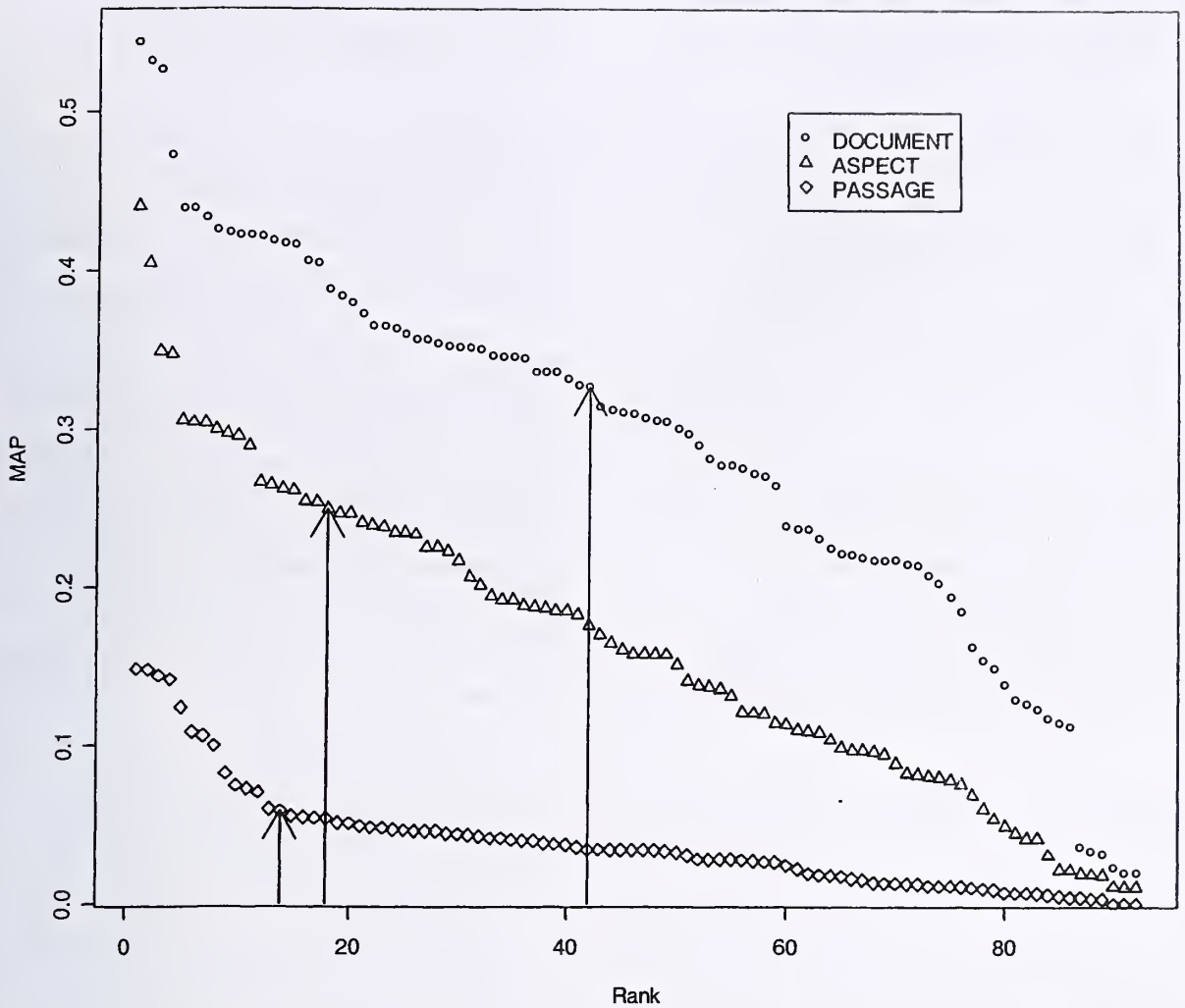


Figure 4 - Ranked MAP showing the first run that is statistically significantly different from the best run.

8. Analysis

Overall there was a wide variation in system performance across submissions for each of the three measures. In general, scores grouped into three sets. A few groups dominated the high scores of each measure, followed by a large group with scores around the mean, and then another large group of relatively low scores. Submissions that scored well on document retrieval tended to score well on both passage and aspect. While a correlation between passage and document retrieval might have been expected, the correlation between document and aspect retrieval is more surprising since aspect retrieval places an emphasis on novelty and document retrieval does not.

Certainly the task and the three measures provided a significant challenge to the participants. The best scores for document retrieval were moderate, and the highest scores on the passage and aspect measures were moderate and fairly low, respectively. No MAP for any system or measure was much greater than 0.50.

For all three measures, the best automatic approaches were as good or better than manual or interactive systems. Manual and interactive approaches did not appear to provide an advantage over automatic methods. However, because the definitions of automatic, manual, and interactive were not as solid as in previous

years because systems had the topic questions available during development, inference should be limited from these observations.

Although a comprehensive analysis was not performed, it was clear from the results and techniques of the top-performing groups in passage retrieval that certain approaches were quite effective. In particular, "trimming" passages to shorten them was done in all the runs with the highest passage MAP. Indeed, some groups noted that non-content manipulations of passages had substantial effects on passage MAP, with one group claiming that breaking passages in half with no other changes doubled their (otherwise low) score. To this end, we defined an alternative passage MAP (PASSAGE2) that calculated MAP as if each character in each passage were a ranked document. In essence, the output of passages was concatenated, with each character being from a relevant passage or not. The complete results are shown in Appendix 6, and summarized in Figure 5, where it can be seen that some re-ranking of runs occurred.

9. Conclusions and Future Directions

This novel approach to the TREC 2006 Genomics Track was carried out successfully, leading to the development of a new test collection with new documents and tasks, as well as a new evaluation method and the software to administer and score it. While further analysis of results is required for more definitive conclusions, it can be noted that passage retrieval in this context is quite difficult, with results quite low. Fortunately, retrieval at the aspect and document levels is much better, indicating users still might be able to find answers to their questions in the biomedical literature. Duplicate relevance assessments showed relatively good levels of reproducibility, with one exceptional outlier.

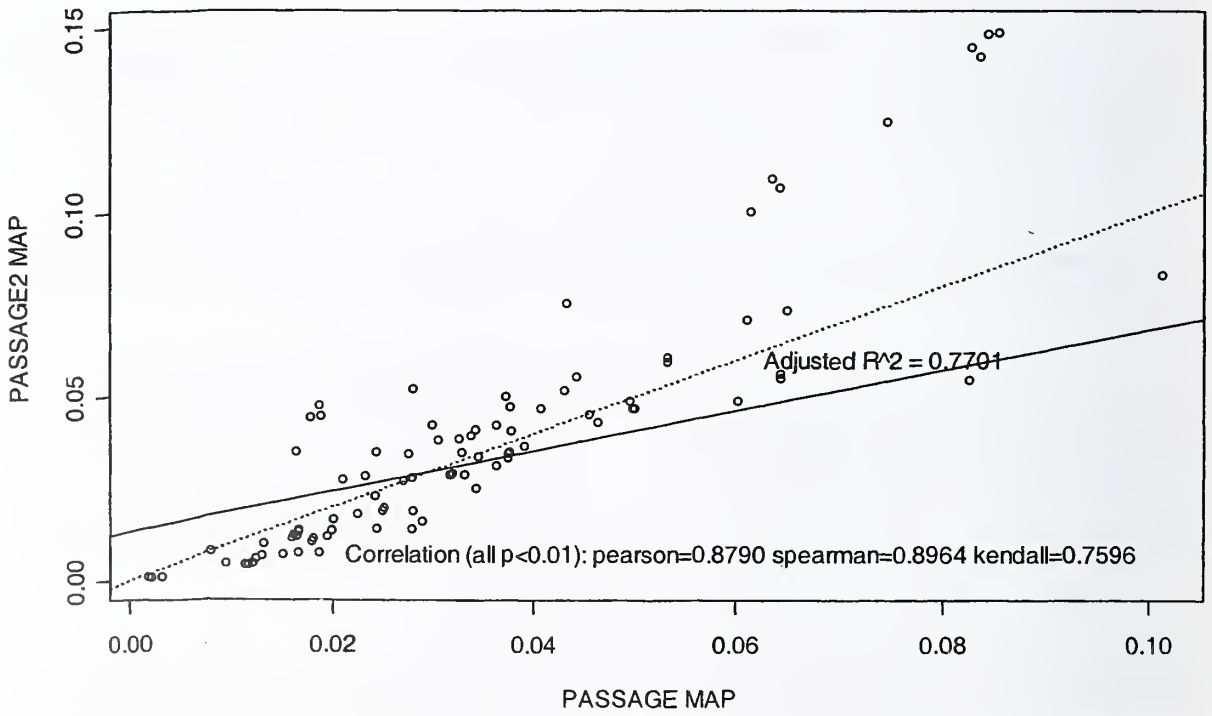


Figure 5 - Comparison of runs using original passage MAP and revised measure (PASSAGE2).

We plan to continue the TREC 2007 Genomics Track in the same direction, using the existing document collection and task structure but adding completely new topics and potentially new topic types. The 2007 track will be the last running of the Genomics Track within TREC, although future options to continue biomedical IR challenge evaluations are being explored.

Acknowledgements

The TREC Genomics Track is funded by grant ITR-0325160 from the U.S. National Science Foundation. The track also thanks Ellen Voorhees, Ian Soboroff, and Lori Buckland of NIST for their help in various ways. We also thank John Sack and Highwire Press (www.highwire.org) for facilitating the use of documents from the respective publishers.

References

Allan, J. (2003). HARD Track overview in TREC 2003 - high accuracy retrieval from documents. *The Twelfth Text REtrieval Conference - TREC 2003*, Gaithersburg, MD. Naitonal Institute of Standards and Technology. 24-37.

Allan, J. (2004). HARD Track overview in TREC 2004 - high accuracy retrieval from documents. *The Thirteenth Text Retrieval Conference (TREC 2004)*, Gaithersburg, MD. National Institute of Standards and Technology.

Cohen, A. and Hersh, W. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics,* 6: 57-71.

Hersh, W. (2001). Interactivity at the Text Retrieval Conference (TREC). *Information Processing and Management,* 37: 365-366.

Hersh, W. (2005). Evaluation of biomedical text mining systems: lessons learned from information retrieval. *Briefings in Bioinformatics,* 6: 344-356.

Hersh, W., Bhupatiraju, R., et al. (2006). Enhancing access to the bibliome: the TREC 2004 Genomics Track. *Journal of Biomedical Discovery and Collaboration,* 1: 3.

Hersh, W., Cohen, A., et al. (2005). TREC 2005 Genomics Track overview. *The Fourteenth Text Retrieval Conference - TREC 2005*, Gaithersburg, MD. National Institute for Standards & Technology.

McGinnis, S. and Madden, T. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research,* 32: W20-W25.

Roberts, P. (2006). Mining literature for systems biology. *Briefings in Bioinformatics,* 7: 399-406.

Appendices

Appendix 1 - List of journals in TREC 2006 Genomics Track full-text collection.

Journal Name	File Size (MB)	Number of Docs	Journal URL
American Journal of Epidemiology	24	1777	aje.oxfordjournals.org
American Journal of Physiology - Cell Physiology	62	2906	ajpcell.physiology.org
American Journal of Physiology - Endocrinology And Metabolism	48	2462	ajpendo.physiology.org
American Journal of Physiology - Gastrointestinal and Liver Physiology	48	2472	ajpgi.physiology.org
American Journal of Physiology - Heart and Circulatory Physiology	99	5170	ajpheart.physiology.org
American Journal of Physiology - Lung Cellular and Molecular Physiology	48	2426	ajplung.physiology.org
American Journal of Physiology - Renal Physiology	39	1897	ajprenal.physiology.org
Alcohol and Alcoholism	9.7	657	alcalc.oxfordjournals.org
Journal of Andrology	7.1	482	www.andrologyjournal.org
Annals of Oncology	16	1273	annonc.oxfordjournals.org
British Journal of Anaesthesia	21	1843	bja.oxfordjournals.org
The British Journal of Psychiatry	17	1531	bjp.rcpsych.org
Blood	209	11291	www.bloodjournal.org
Carcinogenesis	36	2022	carcin.oxfordjournals.org
Cerebral Cortex	22	917	cercor.oxfordjournals.org
Development	62	2402	dev.biologists.org
Diabetes	37	2156	diabetes.diabetesjournals.org
Endocrinology	104	5517	endo.endojournals.org
European Heart Journal	15	1160	eurheartj.oxfordjournals.org
Glycobiology	15	719	glycob.oxfordjournals.org
Human Reproduction	50	3784	humrep.oxfordjournals.org
Human Molecular Genetics	58	3105	hmg.oxfordjournals.org
International Journal of Epidemiology	13	1203	ije.oxfordjournals.org
International Immunology	23	1175	intimm.oxfordjournals.org
Journal of Antimicrobial Chemotherapy	29	2720	jac.oxfordjournals.org
Journal of Applied Physiology	105	5751	jap.physiology.org
Journal of Biological Chemistry	74	4368	www.jbc.org
Journal of Biological Chemistry	33	4733	www.jbc.org
Journal of Biological Chemistry	60	3098	www.jbc.org
Journal of Biological Chemistry	59	2918	www.jbc.org
Journal of Biological Chemistry	49	2432	www.jbc.org
Journal of Biological Chemistry	111	5361	www.jbc.org
Journal of Biological Chemistry	69	3262	www.jbc.org
Journal of Biological Chemistry	119	5539	www.jbc.org
Journal of Biological Chemistry	76	3510	www.jbc.org
Journal of Biological Chemistry	132	6214	www.jbc.org
Journal of Biological Chemistry	109	4886	www.jbc.org
The Journal of Cell Biology	93	3996	www.jcb.org

Journal of Clinical Endocrinology & Metabolism	6.9	758	jcem.endojournals.org
Journal of Cell Science	54	2417	jcs.biologists.org
Journal of Experimental Biology	41	1911	jeb.biologists.org
Journal of Experimental Medicine	70	3492	www.jem.org
The Journal of General Physiology	25	1014	www.jgp.org
Journal of General Virology	40	2375	vir.sgmjournals.org
Journal of Histochemistry and Cytochemistry	24	1592	www.jhc.org
Journal of the National Cancer Institute	34	3214	jncicancerspectrum.oxfordjournals.org
Journal of Neurophysiology	68	2874	jn.physiology.org
Molecular & Cellular Proteomics	9.5	426	www.mcponline.org
Microbiology	46	2400	mic.sgmjournals.org
Molecular Biology and Evolution	25	1303	mbe.oxfordjournals.org
Molecular Endocrinology	36	1610	mend.endojournals.org
Molecular Human Reproduction	14	817	molehr.oxfordjournals.org
Nucleic Acids Research	126	7606	nar.oxfordjournals.org
Nephrology Dialysis Transplantation	38	3629	ndt.oxfordjournals.org
Protein Engineering Design and Selection	15	834	peds.oxfordjournals.org
Physiological Genomics	13	656	physiolgenomics.physiology.org
Rheumatology	21	1985	rheumatology.oxfordjournals.org
RNA	11	544	www.rnajournal.org
Toxicological Sciences	33	1667	toxsci.oxfordjournals.org

Appendix 2 - Relevance judgments per part or all of each maximum-length legal span sent to the judge. If any part of the span was judged relevant, it was counted as definitely or possibly relevant in this table; otherwise it was counted as not relevant.

Topic	Definitely Relevant	Possibly Relevant	Not Relevant	Total
160	214	179	607	1000
161	40	28	932	1000
162	1	17	982	1000
163	99	163	738	1000
164	4	3	993	1000
165	7	10	983	1000
166	2	32	966	1000
167	158	50	792	1000
168	56	187	757	1000
169	54	48	898	1000
170	28	8	964	1000
171	14	36	950	1000
172	305	46	648	999
173	0	0	1000	1000
174	18	18	964	1000
175	0	33	967	1000
176	4	10	986	1000
177	6	3	991	1000
178	3	4	993	1000
179	1	12	987	1000
180	0	0	1000	1000
181	418	162	420	1000
182	94	50	856	1000
183	0	19	981	1000
184	3	2	995	1000
185	17	8	975	1000
186	281	107	612	1000
187	1	2	997	1000

Appendix 3 - Number, average length, and standard deviation of relevant passages and number of aspects per topic.

Topic	Number of Relevant Passages	Mean Passage Length	Standard Deviation of Passage Length	Number of Distinct MeSH Aspects
160	527	307	234	32
161	68	390	449	94
162	18	350	334	20
163	262	289	171	35
164	7	405	210	14
165	17	251	125	11
166	34	485	553	19
167	208	605	612	35
168	243	251	186	35
169	103	1012	1077	32
170	36	234	168	23
171	50	306	134	13
172	593	171	232	78
173	0	0	0	0
174	36	461	285	12
175	33	416	554	27
176	14	412	281	9
177	9	366	240	12
178	7	410	155	21
179	13	360	283	7
180	0	0	0	0
181	589	775	691	96
182	144	293	239	35
183	19	188	116	11
184	5	318	103	10
185	25	209	183	55
186	388	286	291	32
187	3	1107	954	13
Total	3451			781

Appendix 4 - List of groups, runs, run type, and a brief description (provided by the group) for the TREC 2006 Genomics Track

Group	Run	Type	Brief Description
arizona-stateu.gonzalez	asubaral	automatic	First complete run after question variants in.
arizona-stateu.gonzalez	asubaral2	automatic	Using subject-verb-object as part of ranking together with keyword frequency, distance between keywords.
arizona-stateu.gonzalez	asubaral3	automatic	Similar to first run, but less restrictive in filtering. Only require the subject to be in the passage.
concordiau.bergler	BioKI1	interactive	Weighted keyphrases interactively optimized over 2005 data for each query. Output limited to sentence boundaries.
concordiau.bergler	BioKI2	interactive	Weighted keyphrases interactively optimized over 2005 data for each query. Output limited to paragraph boundaries.
concordiau.bergler	BioKI3	interactive	Weighted keyphrases (weight fixed at 25) interactively optimized over 2005 data for all queries. Output limited to paragraph boundaries.
dalianu.yang	DUTgen1	interactive	Rocchio feedback based on 2005's gold standard, Two levels of indexes, BM25, Paragraph-first reranking
dalianu.yang	DUTgen2	interactive	Rocchio feedback based on 2005's gold standard, Two levels of indexes, BM25, Combining reranking
dalianu.yang	DUTgen3	interactive	Rocchio feedback and SVM based on 2005's gold standard, Two levels of indexes, BM25, Paragraph-first reranking
erasmus.schuemie	EMCUT1	automatic	Document retrieval is performed using a language-modelling approach. Passage selection is based on identification of concepts from the UMLS metathesaurus and a gene thesaurus in both the query and the documents.
erasmus.schuemie	EMCUT2	manual	Document retrieval is performed using a language-modelling approach. Passage selection is based on identification of concepts from the UMLS metathesaurus and a gene thesaurus in both the query and the documents. The concepts identified in the query were manually checked and corrected.
fudanu.niu	fdugen1	manual	passage retrieval, svm classification.
fudanu.niu	fdugen2	manual	passage retireval , svm classification, less positive files
fudanu.niu	fdugen3	manual	sentence retrieval, pattern matching.
iit.urbain	iitx1	automatic	sentMatchRatioNormSC + passMatchRatioNormSC
iit.urbain	iitx2	automatic	sentmatchrationormsc+sentnormsc+passmatchrationormsc+passnormsc)/4
iit.urbain	iitx3	automatic	(1*sentmatchrationormsc+0.1*passmatchrationormsc+0.01*sentnormsc+0.001 *passnormsc)
inst-infocomm-res.yu	i2rg061	automatic	document retrieval
inst-infocomm-res.yu	i2rg062	automatic	document reranking
inst-infocomm-res.yu	i2rg063	automatic	Passage Retrieval
kyotou.wan	kyoto1	automatic	Paragraph-level IR with impact-based retrieval and a probabilistic model for term co-occurrence with their scores merged. Queries expanded automatically with synonyms.
kyotou.wan	kyoto2	automatic	a combination of IR impact-based retrieval at document level with a probabilistic model of term coocurance at paragraph level; for the first phase, queries are automatically expanded using synonyms.
kyotou.wan	kyoto20	automatic	a combination of IR impact-based retrieval at document level with a probabilistic model of term coocurance at paragraph level; for the first phase, original queries are employed.
nlm.aronson	NLMfusion	automatic	This run is the equally-weighted fusion of the results of four automatic methods (1) Essie, a search engine developed specifically for biomedical text supporting flexible query expansion; (2) NCBI, a method that performs selective query expansion based on theme analysis; (3) UniGe, a method based on the EasyIR search engine using term and document weightings as well as pivoted normalization; and (4) Smart, a method based on the Smart search engine. Automatic query expansion based on MetaMap and Theme was available to each of the basic methods. Each method produced paragraphs which were then merged into a final list.
nlm.aronson	NLMinter	interactive	This run consists of manually constructed queries generally consisting of a conjunction of topic terms each of which is a disjunction of synonyms. The synonyms were obtained both by introspection and by consulting databases such as Entrez Gene, GeneCards and MeSH. Query development sometimes also involved examination of PubMed and Essie results of preliminary query

			formulations. The queries were processed by Essie, and the results were automatically trimmed of text unrelated to the topics.
nlm.aronson	NLMmanual	manual	This is similar to the automatic Essie method which is part of our automatic fusion run but with some manually modified queries and with results automatically trimmed of text unrelated to the topics.
ntu.chen	NTUadh1	automatic	The underlying retrieval model is KL-divergence. Synonyms for query expansion are selected by checking that the synonyms co-occur with the original query terms in Pubmed's Medline abstracts.
ntu.chen	NTUadh2	automatic	A baseline run using KL-divergence retrieval model.
ntu.chen	NTUadh3	manual	Same as NTUadh1, except that Nur-77 is manually added to queries containing Nurr-77.
ohsu.hersh	OHSUBigclu	automatic	Same as cluster run. Reranking by clustering of similar returns. Parameters for clustering were modified so that cluster were looser.
ohsu.hersh	OHSUCluster	automatic	Same as noclu. The returned passages were further processed by clustering with CLUTO. Features for clustering are text words from the passage with stopwords filtered out and stemming.
ohsu.hersh	OHSUNoclu	automatic	Automatically generated queries with concept expansion. Documents indexed at legal span granularity with Lucene. Retrieved passages scored by tfidf.
purdueu.si	PCPsgAspect	automatic	Combine multiple types of resources for constructing queries; Hierarchical language model smoothing; Post result filter; Aspect retrieval based on vector representation of MMR
purdueu.si	PCPsgClean	automatic	Combine multiple types of resources for constructing queries; Hierarchical language model smoothing; Post result filter
purdueu.si	PCPsgRescore	automatic	Combine multiple types of resources for constructing queries; Hierarchical language model smoothing; Post result filter; Combine multiple types of evidence
queenslandu.geva	Baseline1M	automatic	Baseline run, Identify paragraphs
queenslandu.geva	Z1KL5KX	automatic	Legal span
queenslandu.geva	Z1KL5KY	automatic	Max 5K span
queenslandu.geva	zoom0p5K1M	automatic	Identify complete paragraphs
queenslandu.geva	zoom1K1M	automatic	zoom on passage (500 chars either size)
suny-buffalo.ruiz	UBexp1	automatic	This run uses a pre-retrieval query expansion method that adds gene names and synonyms. Retrieval is performed using SMART Lnu.ltu and returning full paragraphs.
suny-buffalo.ruiz	UBexp1M	automatic	The run has been generated with SMART using pivoted normalization.
suny-buffalo.ruiz	UBexp2	automatic	This run uses automatic pre-retrieval query that adds gene names and synonyms. Retrieval is performed using SMART with atn ann weighting scheme. Retrieval step returns full paragraphs.
suny-buffalo.ruiz	UBexp2M	automatic	The run has been generated with SMART using pivoted normalization (2nd run from Miguel Ruiz).
technion.gabrilovich	LARAg06pe0	automatic	In the preprocessing phase, documents are indexed with BOW and with an additional set of knowledge-rich features based on Wikipedia concepts. First, a simple BOW query is generated from the topic (no expansion or other enhancements). Then, the top 10 returned documents are mapped into most relevant Wikipedia concepts. The resulting concepts are used to query the second index of documents. No explicit domain-specific knowledge is used. Due to lack of time, retrieval is of entire paragraphs, not passages.
technion.gabrilovich	LARAg06pe5	automatic	Note this run is identical to LARAg06pe0 except the use of query expansion. In the preprocessing phase, documents are indexed with BOW and with an additional set of knowledge-rich features based on Wikipedia concepts. First, a simple BOW query is generated from the topic, with blind feedback query expansion. Then, the top 10 returned documents are mapped into most relevant Wikipedia concepts. The resulting concepts are used to query the second index of documents. No explicit domain-specific knowledge is used. Due to lack of time, retrieval is of entire paragraphs, not passages.
technion.gabrilovich	LARAg06t	automatic	Document and query are represented using features generated by an auxiliary classifier that was built using world knowledge extracted from Wikipedia. No other domain-specific or general information is used. Due to lack of time, retrieval is of entire paragraphs, not passages.
tsinghuau.zhang	THU1	automatic	Our best result.
tsinghuau.zhang	THU2	automatic	Shorter Passages to return.
tsinghuau.zhang	THU3	automatic	Longer Passages to return.

uamsterdam.meij	UAmsBaseLine	automatic	Baseline. Just some naive index-specific acronym axpansion on identified (and extracted) NP's
uamsterdam.meij	UAmsExp	automatic	Massive query expansion, using online resources and iteratively gathered acronyms
uamsterdam.meij	UAmsExpSel	automatic	Automatically identified obligatory terms (and expansions)
ucal-berkeley.larso	biotext1	automatic	Basic run. Returns complete legal spans. Ranking based on Lucene score.
ucal-berkeley.larso	biotext3	automatic	Ranked
ucal-berkeley.larso	biotextweb	automatic	Reranking of the first submission run, using n-grams from the Web.
ucolorado.cohen	uchsc1	interactive	Expanded queries are sent to the search engine Lemur. Results undergo zone filtering, and top remaining Lemur results are sent to a singular value decomposition algorithm to expand the results pool by selecting similar paragraphs based on a latent semantic Dirichlet similarity score. Results of the SVD are filtered using Naive Bayes with lexical and conceptual features with training data dervied from manual evaluation of Lemer output.
ucolorado.cohen	uchsc2	interactive	Expanded queries are sent to the search engine Lemur. Results undergo zone filtering. A second, less strict, set of queries is sent to the Lemur search engine and results are filtered using zone filtering and Naive Bayes with lexical and conceptual features with training data dervied from manual evaluation of Lemer output.
ucolorado.cohen	uchsc3	manual	Expanded queries are sent to the search engine Lemur. Results undergo zone filtering.
uguelph.song	UofG0	automatic	Retrieval based on the language modeling approach.
uguelph.song	UofG1	automatic	Retrieval based on the language modeling approach. The results are further filtered based on document coverage.
uguelph.song	UofG2	automatic	Retrieval based on language modeling approach. The results are further filtered based on document and aspect coverage.
uhosp-geneva.ruch	UniGe	automatic	Use the easyIR engine a vector-space with tf.idf weightings and a modified version of pivoted normalization. Basic run.
uhosp-geneva.ruch	UnigeGO	automatic	Use the easyIR engine a vector-space with tf.idf weightings and a modified version of pivoted normalization. GO specific reranking.
uhosp-geneva.ruch	UnigeMesh	automatic	Use the easyIR engine a vector-space with tf.idf weightings and a modified version of pivoted normalization. Template-specific semantic filtering and expansion.
uillinois.chicago.zhou	UICGenRun1	automatic	two-dimensional ranking
uillinois.chicago.zhou	UICGenRun2	automatic	two-dimensional ranking query expansion
uillinois.chicago.zhou	UICGenRun3	automatic	2-dimensional ranking; query expansion; passage retrieval
uiowa.eichmann	UIowa06Geno1	automatic	NLP processing of question, entire paragraph returned as result
uiowa.eichmann	UIowa06Geno2	automatic	NLP processing of question, paragraphs contracted to only those sentences mentioning query terms.
uiowa.eichmann	UIowa06Geno3	automatic	NLP processing of question, entire paragraphs returned, but only those at least 300 characters long (as an ad hoc citation exclusion mechanism).
uiuc.zhai	UIUCauto	automatic	Automatic run.
uiuc.zhai	UIUCinter	interactive	Interactive run.
uiuc.zhai	UIUCinter2	interactive	Interactive run 2.
umass.allan	UMassCIIR1	interactive	Query-biased pseudo relevance feedback. 250 word passages with overlap removed.
umass.allan	UMassCIIR1L	interactive	Query-biased pseudo relevance feedback. The UMassCIIR1 run was "legalized" to only be spans from the legalspans file. Legal spans less than 750 chars were excluded.
umass.allan	UMassCIIR2	interactive	Query-biased pseudo relevance feedback. 500 word passages with overlap removed.
uneuchatel.savoy	UniNE1	automatic	Data fusion of two IR systems (based on normalized RSV values using Z-score) IR system 1 Divergence from randomness, word-based indexing, spelling correction & word variant generation IR system 2 Divergence from randomness, 5-gram indexing
uneuchatel.savoy	UniNE2	automatic	Data fusion of two IR systems (based on normalized RSV values (max)) IR system 1 Divergence from randomness, word-based indexing, spelling correction & word variant generation the document title is included to all passages generated from the article IR system 2 Divergence from randomness, 5-gram indexing
uneuchatel.savoy	UniNE3	automatic	Data fusion of two IR systems (based on normalized RSV values (Z-score), baserun for comparisons) IR system 1 Divergence from randomness, word-

			based indexing IR system 2 Divergence from randomness, 5-gram indexing
utokyo.ishii	Tlab6rGT1	automatic	Automatically calculating abstract level of biomedical concepts and disambiguation of them.
utokyo.ishii	Tlab6rGT2	automatic	Automatically calculating abstract level of biomedical concepts and disambiguation of them. Another condition.
utokyo.ishii	Tlab6rGT3	automatic	Automatically calculating abstract level of biomedical concepts and disambiguation of them. Yet another condition.
uwisconsin.madison	WiscRun1	automatic	Performs POS chunking on topic questions to identify significant noun phrases - Automatically generates expansion term lists for each NP using the MeSH database - Uses Lemur/Indri toolkit to execute queries that require one item in each term list to be found in a paragraph - Ranks results using likelihood of paragraphs given all the expansion term lists concatenated together - Adjusts passage boundaries to include only sentences between the first and last occurrence of key terms
uwisconsin. madison	WiscRun2	automatic	Begins with the same baseline results as our WiscRun1 run - Re-Ranks these results by performing hierarchical clustering on passage bag-of-words vectors - Interleaves results from clusters to promote aspect diversity (Note that clusters are repeatedly considered in order of their average initial rank)
uwisconsin. madison	WiscRun3	automatic	same baseline results as our WiscRun1 run - Re-Ranks using GRASSHOPPER, a graph theoretical algorithm that Performs random walk with absorbing states on the results, to Automatically balance the representativeness and diversity of the final rank
weill-med-cornellu	icb1	interactive	Run 1 was performed with queries at the full article level only. Slider position 200. In this run, we used the MG4J Vigna scorer as baseline. The Vigna scorer favors matches where search terms appear in short text intervals. All runs are performed with the Twease slider at position 200. At this position, the slider expands the query with all the morphological word variants, abbreviations, and MeSH synonyms that match the query words. Morphological word variants are discovered at runtime, with a statistical model trained on Medline 2006 (Campagne, F. unpublished, 2006). Passages are assigned as the minimal intervals where the query match the documents.
weill-med-cornellu	icb2	interactive	Run 2 was performed with parts of the queries at the sentence-level, when appropriate, other terms matching the rest of the article, and ranking by context. Slider at position 200. Context ranking is a new ranking strategy implemented in our textractor framework for the 2006 TREC genomics track. Context queries are expressed as (query)/(context). Briefly, context ranking allows to rank documents matching query by a context, specified as a query expression (e.g., "colon cancer" as a phrase or keywords with boolean clauses). The words in the context do not necessarily occur in the document being ranked. The documents matching the context part of the query are used to infer words that are associated with the context in the corpus. These words are then used to rank the specific set of documents. All runs are performed with the Twease slider at position 200. At this position, the slider expands the query with all the morphological word variants, abbreviations, and MeSH synonyms that match the query words. Morphological word variants are discovered at runtime, with a statistical model trained on Medline 2006 (Campagne, F. unpublished, 2006). Passages are assigned as the minimal intervals where the query match the documents.
weill-med-cornellu	icb3	interactive	Run 3 was performed with queries at the full article level, ranked by context as in Run 2. The context of queries in Run 2 were added to queries from Run 1 to form queries for this run. Slider at position 200. For each topic, queries have the form (query run 1) / (context run 2). All runs are performed with the Twease slider at position 200. At this position, the slider expands the query with all the morphological word variants, abbreviations, and MeSH synonyms that match the query words. Morphological word variants are discovered at runtime, with a statistical model trained on Medline 2006 (Campagne, F. unpublished, 2006). Passages are assigned as the minimal intervals where the query match the documents.
yorku.huang	york06ga1	automatic	1. Use Okapi BM25 for concept-based structured query 2. Use the blind feedback with term selection technique 3. Use a dual index model for passage retrieval 4. No aspect-level retrieval
yorku.huang	york06ga3	automatic	Split the top 500 retrieved passages into 5 groups with 100 passages in each

			group and then use the EM clustering algorithm to re-rank the 100 passages in each group for aspect-level retrieval
yorku.huang	york06ga4	automatic	This run is for document-level retrieval. That is documents will appear in the front of list for only once and those retrieved by different passage previously will be put at the end of list. No aspect-level retrieval.

Appendix 5 - Results of runs sorted by passage, aspect, and document MAP.

Run	Passage MAP	Run	Aspect MAP	Run	Document MAP
THU2	0.1486	UICGenRun1	0.4411	UICGenRun1	0.5439
UICGenRun3	0.1479	NLMinter	0.4051	UICGenRun3	0.532
THU1	0.1442	UICGenRun3	0.3492	UICGenRun2	0.5269
THU3	0.1419	UICGenRun2	0.3479	NLMinter	0.473
UICGenRun2	0.1244	THU1	0.3058	THU1	0.4395
PCPsgRescore	0.1088	THU3	0.3047	THU3	0.4395
PCPsgAspect	0.1065	THU2	0.304	THU2	0.4335
PCPsgClean	0.0999	PCPsgAspect	0.2997	iitx1	0.4261
NLMinter	0.0827	UIUCinter	0.2976	UIUCinter2	0.4243
UICGenRun1	0.075	PCPsgRescore	0.2958	PCPsgRescore	0.4228
DUTgen2	0.073	UIUCinter2	0.29	PCPsgClean	0.4223
DUTgen1	0.0707	NLMmanual	0.2664	PCPsgAspect	0.4217
UIUCinter2	0.0604	PCPsgClean	0.2652	uchsc2	0.4189
UIUCinter	0.0591	iitx1	0.2624	UIUCinter	0.4176
uchsc2	0.056	NLMfusion	0.2617	iitx3	0.4161
iitx1	0.0549	iitx3	0.2546	uchsc1	0.4066
uchsc1	0.0546	BioKI2	0.2537	uchsc3	0.4042
uchsc3	0.0542	uchsc1	0.2496	iitx2	0.3885
icb1	0.0517	uchsc2	0.2472	UIUCauto	0.3842
iitx3	0.0513	uchsc3	0.2467	NLMfusion	0.3793
UofG0	0.0496	UIUCauto	0.2407	UniNE3	0.3725
UIUCauto	0.0486	biotext1	0.2397	UofG1	0.3655
UAmsExpSel	0.0484	Tlab6r2GT3	0.2386	NLMmanual	0.3648
i2rg061	0.0473	Tlab6r2GT2	0.2351	DUTgen1	0.3634
NLMmanual	0.047	NTUadh2	0.2349	DUTgen2	0.3601
NLMfusion	0.0466	Tlab6rGT1	0.2338	NTUadh3	0.3571
NTUadh1	0.0465	UniNE3	0.2259	NTUadh1	0.3563
NTUadh3	0.0464	NTUadh1	0.2256	UniNE1	0.3539
DUTgen3	0.0447	NTUadh3	0.2232	UofG2	0.3526
i2rg063	0.0445	BioKI1	0.2171	biotext1	0.3517
i2rg062	0.0441	UniNE1	0.207	UofG0	0.3517
NTUadh2	0.0429	UniNE2	0.2018	NTUadh2	0.351
BioKI1	0.0419	OHSUNoclu	0.1946	UniNE2	0.346
OHSUNoclu	0.0419	UBexp2	0.1922	EMCUT1	0.3459
UniNE3	0.0407	UBexp2M	0.1922	EMCUT2	0.3459
UBexp2	0.0403	OHSUBigclu	0.1892	york06ga4	0.3444
UBexp2M	0.0403	OHSUCluster	0.188	york06ga1	0.3365
UniNE1	0.039	iitx2	0.1869	UBexp2	0.3364
UniNE2	0.0384	DUTgen1	0.1857	UBexp2M	0.3364
OHSUBigclu	0.0379	UofG0	0.1856	UMassCIIR2	0.3317
iitx2	0.0363	BioKI3	0.1828	OHSUNoclu	0.3274
biotext1	0.0348	UMassCIIR2	0.1761	york06ga3	0.3269
icb2	0.0348	UniGe	0.1702	Tlab6r2GT2	0.3139
BioKI2	0.0346	DUTgen2	0.1648	Tlab6r2GT3	0.3121
UBexp1	0.0346	UofG1	0.1608	Tlab6rGT1	0.3105
UBexp1M	0.0346	UofG2	0.1583	BioKI2	0.3093
OHSUCluster	0.0344	UBexp1	0.1578	BioKI1	0.3072

UniGe	0.0343	UBexp1M	0.1578	OHSUBigclu	0.3051
BioKI3	0.0335	UnigeMesh	0.1577	OHSUCluster	0.3042
UnigeMesh	0.0328	WiscRun1	0.1516	icb1	0.3003
UnigeGO	0.0309	WiscRun3	0.1411	UMassCIIR1	0.2964
Tlab6r2GT2	0.0288	UnigeGO	0.1386	DUTgen3	0.2902
Tlab6r2GT3	0.0287	DUTgen3	0.1379	UnigeMesh	0.2814
Tlab6rGT1	0.0286	UMassCIIR1	0.1361	UBexp1	0.277
UAmsExp	0.0286	WiscRun2	0.1319	UBexp1M	0.277
UofG1	0.0282	kyoto1	0.1217	UniGe	0.2755
Z1KL5KY	0.0277	Z1KL5KX	0.1209	BioKI3	0.2724
UofG2	0.0271	Z1KL5KY	0.1207	UnigeGO	0.2706
Z1KL5KX	0.027	UMassCIIR1L	0.1143	UMassCIIR1L	0.2647
kyoto1	0.0248	UAmsExpSel	0.1137	Z1KL5KY	0.2386
UAmsBaseLine	0.0226	icb1	0.11	Z1KL5KX	0.2375
york06ga1	0.0197	Baseline1M	0.1097	WiscRun1	0.2368
WiscRun1	0.0188	york06ga1	0.1084	UAmsExpSel	0.2312
york06ga3	0.0187	york06ga3	0.1039	kyoto1	0.2248
UMassCIIR1L	0.0179	zoom1K1M	0.099	i2rg062	0.2219
UMassCIIR1	0.0164	biotextweb	0.0974	WiscRun3	0.2208
WiscRun3	0.0159	EMCUT1	0.0972	biotextweb	0.2195
fdugen3	0.0138	york06ga4	0.0964	Baseline1M	0.2176
WiscRun2	0.0137	zoom0p5K1M	0.0952	zoom0p5K1M	0.2176
york06ga4	0.0135	EMCUT2	0.0891	zoom1K1M	0.2176
zoom1K1M	0.0132	LARAg06pe0	0.0833	i2rg061	0.2148
zoom0p5K1M	0.0131	LARAg06pe5	0.0818	i2rg063	0.2135
Baseline1M	0.0121	i2rg061	0.0812	UAmsExp	0.2081
biotextweb	0.0118	i2rg063	0.0802	WiscRun2	0.203
EMCUT1	0.0117	icb2	0.0784	fdugen3	0.1943
EMCUT2	0.0113	i2rg062	0.0758	icb2	0.1846
LARAg06pe0	0.0109	kyoto2	0.0692	UAmsBaseLine	0.1624
LARAg06pe5	0.0103	kyoto20	0.061	LARAg06pe0	0.1542
UMassCIIR2	0.0097	fdugen3	0.0544	fdugen1	0.1488
icb3	0.0076	UAmsExp	0.0495	LARAg06pe5	0.1385
fdugen1	0.0075	UAmsBaseLine	0.0457	kyoto2	0.1297
kyoto20	0.0075	biotext3	0.0419	fdugen2	0.1267
kyoto2	0.0071	LARAg06t	0.0418	kyoto20	0.1231
fdugen2	0.0065	icb3	0.0313	biotext3	0.1178
LARAg06t	0.0056	fdugen1	0.022	icb3	0.1147
biotext3	0.0044	UIowa06Geno3	0.0219	LARAg06t	0.1119
UIowa06Geno2	0.0044	UIowa06Geno1	0.0199	asubaral3	0.0365
UIowa06Geno1	0.0039	fdugen2	0.0193	asubaral	0.0334
UIowa06Geno3	0.0039	UIowa06Geno2	0.0187	asubaral2	0.0319
asubaral3	0.0008	asubaral	0.0116	UIowa06Geno1	0.0234
asubaral	0.0007	asubaral2	0.0114	UIowa06Geno2	0.02
asubaral2	0.0007	asubaral3	0.011	UIowa06Geno3	0.0198
Mean	0.0392	Mean	0.1643	Mean	0.2887
Median	0.0345	Median	0.1581	Median	0.3083
Min	0.0007	Min	0.011	Min	0.0198
Max	0.1486	Max	0.4411	Max	0.5439

Appendix 6 - Comparison of results and ranks of original (PASSAGE) and modified (PASSAGE2) passage MAP.

Run	PASSAGE MAP	PASSAGE2 MAP	PASSAGE MAP Rank	PASSAGE2 MAP Rank
THU2	0.148593	0.085316	1	2
UICGenRun3	0.147916	0.084342	2	3
THU1	0.144239	0.082738	3	5
THU3	0.141929	0.083562	4	4
UICGenRun2	0.124390	0.074536	5	7
PCPsgRescore	0.108766	0.063310	6	12
PCPsgAspect	0.106500	0.064048	7	11
PCPsgClean	0.099922	0.061270	8	13
NLMinter	0.082714	0.101262	9	1
UICGenRun1	0.075050	0.043047	10	24
DUTgen2	0.073024	0.064767	11	8
DUTgen1	0.070666	0.061039	12	14
UIUCinter2	0.060380	0.053200	13	16
UIUCinter	0.059062	0.053124	14	17
uchsc2	0.055976	0.064229	15	10
iitx1	0.054941	0.044172	16	23
uchsc1	0.054570	0.064268	17	9
uchsc3	0.054223	0.082599	18	6
icb1	0.051705	0.027911	19	52
iitx3	0.051309	0.042971	20	25
UofG0	0.049608	0.037067	21	35
UIUCauto	0.048644	0.049393	22	20
UAmsExpSel	0.048445	0.060108	23	15
i2rg061	0.047251	0.018594	24	70
NLMmanual	0.047048	0.037467	25	30
NLMfusion	0.046584	0.040631	26	26
NTUadh1	0.046493	0.049792	27	19
NTUadh3	0.046379	0.049894	28	18
DUTgen3	0.044680	0.045511	29	22
i2rg063	0.044458	0.018773	30	68
i2rg062	0.044096	0.017759	31	73
NTUadh2	0.042941	0.046341	32	21
BioKI1	0.041915	0.036084	33	37
OHSUNoclu	0.041866	0.029858	34	49
UniNE3	0.040747	0.034017	35	40
UBexp2	0.040306	0.037583	36	28
UBexp2M	0.040306	0.037583	37	29
UniNE1	0.038983	0.033616	38	41
UniNE2	0.038359	0.032431	39	44
OHSUBigclu	0.037946	0.030458	40	48
iitx2	0.036266	0.039009	41	27
icb2	0.034804	0.016423	42	78
biotext1	0.034778	0.024210	43	61
UBexp1	0.034650	0.037421	44	31
UBexp1M	0.034650	0.037421	45	32
BioKI2	0.034603	0.032756	46	43

OHSUCluster	0.034366	0.027431	47	55
UniGe	0.034294	0.037394	48	33
BioKI3	0.033540	0.034368	49	38
UnigeMesh	0.032847	0.037338	50	34
UnigeGO	0.030936	0.036192	51	36
Tlab6r2GT2	0.028798	0.031839	52	45
Tlab6r2GT3	0.028680	0.031514	53	47
Tlab6rGT1	0.028639	0.031713	54	46
UAmsExp	0.028589	0.033008	55	42
UofG1	0.028192	0.023125	56	62
Z1KL5KY	0.027745	0.027867	57	53
UofG2	0.027139	0.020943	58	64
Z1KL5KX	0.026958	0.026984	59	56
kyoto1	0.024776	0.034174	60	39
UAmsBaseLine	0.022634	0.024233	61	60
york06ga1	0.019689	0.025089	62	57
WiscRun1	0.018783	0.027995	63	51
york06ga3	0.018684	0.024995	64	58
UMassCIIR1L	0.017901	0.022477	65	63
UMassCIIR1	0.016448	0.020021	66	65
WiscRun3	0.015890	0.028893	67	50
fdugen3	0.013789	0.027836	68	54
WiscRun2	0.013735	0.024343	69	59
york06ga4	0.013542	0.019954	70	66
zoom1K1M	0.013236	0.016688	71	75
zoom0p5K1M	0.013069	0.016699	72	74
Baseline1M	0.012056	0.016280	73	79
biotextweb	0.011773	0.019502	74	67
EMCUT1	0.011705	0.016463	75	77
EMCUT2	0.011284	0.016136	76	80
LARAg06pe0	0.010871	0.018109	77	71
LARAg06pe5	0.010268	0.017938	78	72
UMassCIIR2	0.009669	0.013213	79	82
icb3	0.007629	0.008082	80	89
kyoto20	0.007493	0.016605	81	76
fdugen1	0.007480	0.018689	82	69
kyoto2	0.007093	0.015123	83	81
fdugen2	0.006471	0.013105	84	83
LARAg06t	0.005562	0.012587	85	84
biotext3	0.004443	0.009515	86	88
UIowa06Geno2	0.004425	0.012204	87	85
UIowa06Geno1	0.003856	0.011418	88	87
UIowa06Geno3	0.003851	0.011783	89	86
asubaral3	0.000759	0.003154	90	90
asubaral2	0.000690	0.001915	91	92
asubaral	0.000684	0.002142	92	91

TREC-2006 Legal Track Overview

Jason R. Baron, National Archives and Records Administration, Office of General Counsel, Suite 3110, College Park, MD 20740, jason.baron@nara.gov

David D. Lewis, David D. Lewis Consulting, 858 W. Armitage Ave. #296, Chicago, IL 60614, trec06@DavidDLewis.com

Douglas W. Oard, College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, oard@umd.edu

Abstract

This paper describes the first year of a new TREC track focused on "e-discovery" of business records and other materials. A large collection of scanned documents produced by multiple real world discovery requests was adopted as the basis for the test collection. Topic statements were developed using a process representative of current practice in e-discovery applications, with both Boolean and natural language queries being supported. Relevance judgments were performed by personnel who had received professional training, and often considerable experience, in review of similar materials for this task. Six research teams and one manual searcher submitted a total of 33 retrieved sets for each topic. These were pooled and a portion assessed to support evaluation of both the retrieved sets themselves and for future use of the collection.

1. Introduction

The use of information retrieval techniques in law has traditionally focused on providing access to legislation, regulations, and judicial decisions. Searching business records for information pertinent to a case (or "discovery") has also been important, but digitally searchable records were until recently the exception rather than the norm. That is rapidly changing, however. The motivating goal of this new legal track at the Text Retrieval Conference (TREC) is to assess the ability of information retrieval technology to meet the needs of the legal community for tools to help with retrieval of business records. This is an issue of increasing importance given the vast amount of information in electronic form to which access is required during litigation. Ideally, the results of our studies will also help to advance the discussion of the capabilities and limitations of automated support for e-discovery in the legal community.

The importance of doing well at e-discovery is hard to overstate. In the past few years, lawsuits involving giant corporations and single individuals alike have resulted in huge multi-million and even billion dollar adverse verdicts turning on the failure of a party to the litigation to properly preserve and provide access to various forms of electronic records, including most notably e-mail, and data on backup tapes (see, e.g., *Coleman, 2005*; *Zubulake, 2004*). Beyond the headlines, however, are a growing percentage of lawsuits that involve the production of responsive electronic data stored in vast corporate, governmental, and other repositories. Lawyers are struggling to keep up with the profusion of electronic data and metadata in all its forms, on desktops and networks. So too, troves of "legacy" documents, sometimes going back decades, continue to be maintained and need to be searched in response to discovery requests.

The results of the legal track are especially timely and important given recent changes in the U.S. Federal Rules of Civil Procedure that went into effect on December 1, 2006. The amended rules introduce a new category of evidence, namely, "electronically stored information" ("ESI") in "any medium," intended to stand on an equal footing with existing rules covering the production of "documents." Rule 26(f) specifically directs that at an initial conference of the parties, "any issues relating to disclosure or discovery of electronically stored information, including the form or forms in which it should be produced" are to be

discussed. Such issues will necessarily include the need to consider how appropriate ESI is made accessible to opposing parties. Providing access involves more than just search technology, of course—initial query formulation, iterative query refinement, and review of search results for relevance and privilege are important components of the entire process. The Advisory Committee notes to Rule 34 say that in talking about "ESI in any medium," the rules amendments were intended to "encompass future developments in computer technology," which speaks specifically to our goals for the TREC Legal Track.

Against the backdrop of the Federal Rules changes, the status quo in the legal profession, even in large and complex litigation, is continued reliance on free-text Boolean searching for satisfying document (and now ESI) production demands (Sedona Conference 2005). Thus, to the extent a trend exists in the case law, it is where courts have intervened at early stages to ensure that parties negotiate "search protocols." To date these have consisted solely of a static list of agreed upon query terms, rather than more complex forms of negotiations over, for example, complex (extended) Boolean queries (e.g., those specifying truncation and/or proximity operators). Moreover, as of the date of this paper, there is no reported case law in the United States where courts have been called upon to adjudicate the reasonableness of alternative forms of search methodologies (e.g., ranked retrieval). It is only a matter of time, however, before parties in litigation will more fully utilize alternative techniques, enter into negotiations regarding search system selection and/or query formulation, and, inevitably, conduct formal adjudication over the reasonableness and efficacy of such alternative approaches.

An important aspect of e-discovery and thus of the TREC legal track is an emphasis on recall over precision. In light of the fact that a large percentage of requests for production of documents (and now ESI) routinely state that "all" such evidence is to be produced, it becomes incumbent on responding parties to attempt to maximize the number of responsive documents found as the result of a search. All things being equal, lawyers would be expected to move towards alternative search methods that produce greater numbers of responsive documents for the same resources expended; conversely, alternatives that produce fewer responsive documents are likely to be judged as insufficient, even if greater precision (economy) is achieved overall. If recall comparable to the presently used techniques could be assured, then interest would likely exist in increasing precision (thereby diminishing the need to manually review false positive hits generated by automated means).

There have been to date few research efforts studying effectiveness of retrieval in civil discovery contexts. The seminal study (Blair & Maron, 1985), found that while attorneys believed they had found 75% of the relevant documents for litigation involving a train accident, in fact only an estimated 20% of relevant documents were discovered. The authors attributed this to the inherent ambiguity of language. At least one later study has looked at a comparison of Boolean and natural language searches in the context of a structured database of case precedents (Turtle 1994), and experiments with Boolean systems on outside the legal context have been reported at TREC (e.g., Lu et al. 1993; Jacobs 1995) and elsewhere.

The key goal of the TREC 2006 legal track was to apply objective benchmark criteria for comparing search technologies, using topics and documents approximating those of actual discovery settings. Given the reality of the use of Boolean search in present day litigation, of significant interest was comparing the efficacy of Boolean search using negotiated queries with alternative methods. The chosen collection, about seven million scanned documents from the tobacco Master Settlement Agreement, can also be used for technology-centered experiments comparing retrieval techniques based on metadata and/or optical character recognition.

The remainder of this paper is organized as follows. Section 2 describes the document collection. Section 3 then explains the topic development process. In Section 4, the process by which relevance judgments were created is presented. Section 5 identifies the participating research teams and presents some preliminary results. Section 6 concludes the paper.

2. Document Collection

The Legal Track required a collection reflecting the scope and diversity of documents searched in real discovery settings. Obtaining access to the internal documents of large enterprises for research purposes is difficult, but ironically discovery proceedings in real legal cases provide one source of such material. As the Legal Track test collection we chose the IIT CDIP Test Collection, version 1.0 (which we will refer to as "IIT CDIP 1.0") which is based on documents released under the tobacco "Master Settlement Agreement" (MSA).

The MSA settled a range of lawsuits by the Attorneys General of several US states against seven US tobacco organizations (five tobacco companies and two research institutes). One part of this agreement required those organizations to make public on the World Wide Web (through at least June 30, 2010) all documents produced in discovery proceedings in the lawsuits by the states, as well as all documents produced in a number of other smoking and health-related lawsuits. Notable among the provisions is that the tobacco organizations were required to provide to the National Association of Attorneys General (NAAG) a copy of metadata and the scanned documents from the websites, and are forbidden from objecting to any subsequent distribution of this material. The text of the MSA and accompanying appendices and other documents can be found at the websites of Attorneys General of several US states, including California (*http://ag.ca.gov/tobacco/msa.php*).

The University of California San Francisco (UCSF) Library, with support from the American Legacy Foundation, has created a permanent repository, the Legacy Tobacco Documents Library (LTDL), for tobacco documents (Schmidt, Butter & Rider 2002) in order to assure continued availability of these materials. The Illinois Institute of Technology (IIT) Complex Document Information Processing (CDIP) 1.0 collection is based on a snapshot, generated between November 2005 and January 2006, of the MSA subcollection of the LTDL. The snapshot consisted of 1.5 TB of scanned document images, as well as metadata records and optical character recognition (OCR) output produced from the images by UCSF. The IIT CDIP project subsequently reformatted the metadata and OCR, combined the metadata with a slightly different version obtained from UCSF in July 2005, and discarded some documents with formatting problems, to produce the IIT CDIP 1.0 collection (Lewis, et. al 2006).

The IIT CDIP 1.0 collection consists of 6,910,192 document records in the form of XML elements. The two subelements which provide the most conventional target for text retrieval are <ti> (the document title) and <ot> (the OCR text). The highly variable quality of the OCR, combined with the great variations in document length (from one page to thousands of pages) makes retrieval even on these fields a challenge. In addition to the text subelements, there are a wide range of other metadata subelements present in some or all of the records, including senders and recipients, important names mentioned in the document, controlled vocabulary categories, geographical and organizational context identifiers, and many others. The degree to which this information is present varies with the originating tobacco organization and other factors. Overall, the structure of the data is extremely rich and still not well understood.

IIT CDIP 1.0 had strengths and weaknesses as a collection for the Legal Track. The wide range of document genres (including letters, memos, budgets, reports, agendas, minutes, plans, transcripts, scientific articles, email, and many others) and the large number of documents are very typical of legal discovery settings. The fact that documents were scanned and OCR'd is representative of some discovery situations, but perhaps not those of most interest to those concerned with electronic discovery. The rich but variable quality metadata is also perhaps not typical. The fact that the MSA documents were themselves the *output* of legal discovery proceedings might suggest they are unrepresentative as *inputs* to TREC's simulation of a legal discovery situation. Our worries about that point are mitigated to some extent, however, by the fact that the MSA documents originated from seven different organizations in response to hundreds of distinct document requests in multiple legal cases. Thus their diversity is more representative of a diverse population of company records than perhaps might initially be imagined. We further addressed this concern by using a range of topics in the evaluation, some with content highly similar to MSA discovery requests, and others very different. The fact that documents originated from seven different organizations but were searched as a unit is decidedly anomalous from the perspective of federated search, and some future users of the collection may wish to treat the seven subcollections in a more separate manner.

Several minor glitches in the preparation of IIT CDIP 1.0 turned up during indexing of the data by Legal Track participants. In addition, a number of documents turned out to have XML records but no document images, which was both an immediate problem for relevance assessment, and also a problem for the types of document image retrieval and mining studies towards which the CDIP project is targeted (Agam et al. 2006). These problems are being investigated in ongoing work by the IIT CDIP project.

3. Topic Development

Topic development was modeled on U.S. civil discovery practice. In the litigation context, a "Complaint" is filed in court, outlining the theory of the case, including factual assertions and causes of action representing the legal theories of the case. In a regulatory context, often formal letters of inquiry serve a similar purpose by outlining the scope of the proposed investigation. In both situations, soon thereafter one or more parties create and transmit formal "requests for the production of documents" to adversary parties, based on the issues raised in the Complaint or Letter of Inquiry. (If in federal court, this type of demand is typically filed pursuant to Fed. R. Civ. P. 34, but may also be sent to third party non-defendants via subpoena under Fed. R. Civ. P. 45.) Requests to produce documents are typically very broadly worded, in an attempt to force the opposing party to provide a maximum number of responsive documents. In some cases, however, requests are purposely more narrowly tailored when the focus is on particular documents known to be in the possession of a party which are deemed useful at trial. A third category of requests are aimed at finding only particular types of documents (e.g. all "internal memoranda" on a designated topic.)

It is increasingly common for lawyers to consider requesting that specific search terms be used for the purpose of searching large databases for potentially responsive documents. Courts have begun referring to the development of "search protocols," which are to be developed either unilaterally or, to a greater or lesser extent, made subject to negotiations between parties prior to conducting searches. At present, it is typically assumed that an extended Boolean search (i.e., one with truncation and/or proximity operators) will be performed, although some legal technology firms now also support other types of search technology. Less well known is what percentage of cases have utilized a robust or sophisticated process of negotiations over how search terms, wildcards, Boolean logic, and proximity operators are to be combined to form queries. Nevertheless, for the purpose of the TREC 2006 legal track, it was deemed important to develop topics that stood in as proxies for real-life requests to produce documents in which a set of Boolean strings were developed by a negotiation process between two parties.

For the TREC 2006 legal track, five hypothetical complaints were created by members of the Sedona Conference®, a group of lawyers who are leading the development of professional practices for e-discovery. These complaints described: (1) an investigation into a fictional tobacco company's improper campaign contributions; (2) a consumer protection lawsuit challenging a fictional tobacco company's "product placement" decisions in television, film, and theatre shows watched by children; (3) an "insider trading" securities lawsuit involving fictional tobacco executives; (4) an antitrust lawsuit involving the movement of commerce in California; and (5) a product liability lawsuit involving defective surgical devices as shown in animal testing. In using fictional names and jurisdictions, the track coordinators, on behalf of TREC, attempted to ensure that no third party would mistake the academic nature of the TREC legal track for an actual lawsuit against real-world companies, and any would-be link or association with either past or present real litigation involving such companies was entirely unintentional.

For each of the five complaints, a set of topics (formally, "requests to produce") were initially created by the creator of the complaint, and revised by the track co-coordinators. Revisions were considered necessary where the initial topic appeared to have too few or too many relevant documents for effective evaluation, or when it was feared assessors would find the topic too ambiguous. (In this respect, the TREC exercise models real-life objections that often are made to "overbroad," "vague," or "ambiguous" discovery requests, sometimes resulting in courts requiring parties to re-submit narrower and more focused requests.) In the end, 43 topics were selected by the track coordinators for use in the evaluation.

Two aspects of this screening process were less than ideal. First, the evaluation of breadth and ambiguity was done by the track organizers and a professional tobacco searcher, not by the eventual assessor for each

topic, as NIST has often been able to do in past TRECs. (Most assessors had not yet been recruited at the time topics were drawn up.) Second, the screeners did not have access to ranked retrieval search of the collection. Screening was done using the Boolean interface available from UCSF,[1] which at that time had only a beta version of OCR search.

For each of these 43 topics, the initial topic creator and a track coordinator took the roles of requester and respondent (respectively) in a discovery process, and engaged in an iterated negotiation over the form of a Boolean query for the topic. The final XML topic file contained 43 entries, each including the production request, the associated complaint (which for simplicity was repeated in full for each production request associated with that complaint), the extended Boolean query initially proposed by the (simulated) requesting party, the final extended Boolean query that was agreed upon, and any additional extended Boolean queries in the negotiation history. Human-readable versions of the complaints and the production requests were also prepared for use by relevance assessors and interactive searchers, and a cross-reference to each was recorded in the XML topic file. The topic file is available from the track Web page, *http://trec-legal.umiacs.umd.edu.*

4. Relevance Judgments

This section describes the process by which relevance judgments were created.

4.1. Creating Judgment Pools

The complexity of the CDIP documents and topics, and a report of pooling problems with other large collections (Buckley, et al 2006) generated some concern about the adequacy of conventional pooling approaches for the Legal Track. We adopted several strategies for addressing these problems, though none were a complete solution.

We invited track participants to submit up to eight runs (in an effort to maximize pool diversity), asked for runs to depth 5,000 (to facilitate computation of recall-oriented evaluation measures), and asked participating teams to designate their runs for inclusion in the assessment pools in priority order. We included in the assessment pools the top 100 documents from the highest priority run from each team and the top 10 documents from each of the other runs from that team. This yielded a maximum of 170 documents per team for any topic, although usually fewer documents than that were added to the pools because duplicates were removed (both within and across teams). A total of six participating teams submitted a total of 31 runs for official scoring. Two additional runs that were commissioned especially for the track were then used to further enrich the pools.

It is well known that expert searchers can and will often find documents that fully automated term-matching techniques would miss. The IIT CDIP project therefore contracted with an expert tobacco document searcher (Celia White, *http://professionalresearchservices.com*) to produce a set of approximately 100 documents for each topic to add to the pools. Working with a track coordinator, she attempted to find documents that were both relevant to a topic and unlikely to be highly ranked by ranked retrieval systems.

A particular interest in the Legal Track was to compare the effectiveness of the final negotiated Boolean query with the effectiveness of ranked retrieval systems. Hummingbird generously agreed to submit for our use as a baseline Boolean run the retrieved sets resulting from directly executing the negotiated Boolean query (with only a few format corrections, as described in the Open Text[2] team's paper). This run was not counted as an official submission of the Hummingbird team, but rather as a track baseline. We then drew a stratified sample (Cochran, 1977; Lewis, 1996) from the set of documents retrieved by the

[1] http://legacy.library.ucsf.edu/

[2] Hummingbird was acquired by Open Text Corporation in October 2006. Hence the Open Text Corporation paper describes the Hummingbird runs.

negotiated Boolean query for each topic in order to support unbiased estimation of certain evaluation measures for these sets.

Stratification was done by assigning each document from a baseline Boolean set to one of three strata based on whether and how that document occurred in the 31 official submitted runs. The three strata were:

- ***Stratum 1*** (documents occurring in the top 5,000 for at least one official run submitted by each of two or more of the six participating sites),
- ***Stratum 2*** (documents occurring in the top 5,000 for one or more official runs from exactly one of the six participating sites), and
- ***Stratum 3*** (documents not occurring in the top 5,000 for any official run submitted by any participating site).

For each topic, NIST drew a simple random sample of 100 documents from ***Stratum 1***, 50 from ***Stratum 2***, and 50 from ***Stratum 3*** to add to the pool for that topic. When a stratum was exhausted, leftover documents were drawn from the other strata in proportion to their original allocation. Using different stratification strategies for different topics could have improved our estimates, but would have complicated the sampling procedure. An unexpected downside of the above stratification was that ***Stratum 3*** often turned out to be empty. This may have resulted from use of terms from the negotiated Boolean query by ranked retrieval systems, which was allowed (and, indeed, encouraged) by the track guidelines.

One participating team, Hummingbird, leveraged the track's approach to constructing assessment pools (which was known by the participants) to do their own stratified sampling experiment. Their main run (humL06tvz) actually drew documents from various depths of a standard ranked run, enabling them to compute unbiased estimates of precision (and the number of unjudged relevant documents) to depth 9,000. Details can be found in their paper (Tomlinson, 2006). This strategy almost certainly increased the diversity of the assessed pools (at the cost of some richness in relevant documents) by increasing the number of lower-ranked documents assessed. It also invalidated our computation of standard evaluation measures for that run (which are shown in Figure 3 only for completeness).

4.2. Relevance Judgment Process

A total of 35 volunteers from government, law firms, legal technology firms, and law schools (plus two unaffiliated individual volunteers) assessed a total of 32,738 documents in the judgment pools for 40 of the topics. Due to lack of assessment capacity, no assessments were performed for the three remaining topics, and they were thus removed from the evaluation. The volunteers included eight lawyers, ten law students (with 1st, 2nd and 3rd year students all represented), three paralegals with substantial legal experience, one professional archivist, one historian, and several individuals with degrees with science or finance. The affiliations of volunteers for primary assessments were the National Archives and Records Administration (8 topics), George Washington University Law School (D.C., 8 topics), H5 Technologies Inc. (San Francisco, 7 topics), Lewis & Roca LLP (Phoenix, 4 topics), Preston Gates LLP (Seattle, 3 topics), Bank of America (Charlotte, 2 topics), FTI Consulting, (New York City, 2 topics), one topic each by George Mason University School of Law (Virginia), Reasonable Discovery LLC (Virginia), New Mexico State Attorney General's Office, and one topic each from three private individuals (in Florida, California, and the U.K.).

The assessors used a beta version of a Web-based platform to view the scanned MSA documents and record their relevance judgments. (The platform was designed by David D. Lewis Consulting, and implemented by Smokescreen Consulting, as part of the IIT CDIP project.) We provided the assessors with a "How To Guide" (Baron, Lewis & Oard, 2006) that explained that the project was modeled on the ways in which lawyers make and respond to real requests for documents, including in electronic form. Assessors were told to assume that they had been requested by a senior partner, or hired by a law firm or another company, to review a set of documents for "relevance." No special, comprehensive knowledge of the matters discussed in each complaint was expected (e.g., no need to be an expert in federal election law, product liability, etc.). The heart of the exercise was to look for relevant and nonrelevant documents within a topic. Relevance, consistent with all known legal definitions from Wigmore to Wikipedia, was to be

defined broadly. Specifically, assessors were instructed that a document should be considered relevant when the reference to the topic was found in the document. Assessors were reminded that a document may be relevant even if it fails to contain any of the words in the topic request, and conversely, that a document may end up being considered not relevant despite containing one or more words from the topic request. Assessors were also informed that for some topics, the *document type* would circumscribe the scope of the topic (e.g., all *internal memoranda* of a company on topic x), and that (for a very few topics) the scope might be limited by a specified date span (e.g., all documents created in 1992). Relevance judgments were to be recorded as a binary value (yes or no), although a third "unsure" category was also available in the assessment platform.

The first phase of assessment (the only phase initially planned) began on August 7, 2006, and was completed on September 15, 2006. This was the first time that distributed assessment of document images had been used in TREC, and a few complications unsurprisingly arose. It became apparent during assessment that the collection contained some extremely long documents (e.g. a 3,500 page card catalog) and that the participating systems had retrieved a disproportionate number of these long documents. The assessment guidelines were changed in mid-August to allow assessors to mark documents longer than 300 pages as "unsure" if their relevance could not be determined by examining the available metadata and a few pages of the document. Documents marked as unsure were treated as not relevant. When surveyed after completion of their work, some assessors suggested that graded relevance judgments be supported in future years, so as to distinguish between mere "passing references" to a topic (which were recorded as relevant for this year's track) and documents that materially or substantively discuss a topic (which were also recorded as relevant this year).

Some of the assessors went beyond the text of the topic (the complaint, the production request, and the Boolean queries) to perform additional legal research which they viewed as helpful to the exercise. For example, the assessor for Topic 30 researched at greater length what the numbered statutory code provisions were corresponding to the California Cartwright Act, to ensure that all documents containing such references, with or without reference to the Cartwright Act itself, would be marked as responsive. The assessor on Topic 10 performed independent research into the ban on tobacco advertising, as an aid to understanding what documents might be expected to be found in response to a topic involving tobacco product placement in television or film. One assessor asked for assistance on the definition of one of the keywords in the topic, leading to additional research conducted on the Internet.

Some differences were observed in how liberally or narrowly assessors viewed the scope of their discretion to find responsiveness. In some exceptional cases, assessors were willing to find responsiveness even where a key term might be missing, if the document was otherwise sufficiently generic and might yet be viewed as responsive with the aid of further research. For example, the assessor for Topic 9 ("All documents discussing, referencing or relating to payment of compensation to 20th Century Fox Corporation for placement of products and/or brands in a film production") marked certain documents as relevant even if the film company was not expressly mentioned, where the context indicated that the company might be involved. In most cases, however, assessors appeared to adopt relatively restrictive interpretations on what met the mark for relevance.

Assessors reported some confusion as to whether they should exclude documents that might be within the literal scope of a production request when read in isolation, but which weren't relevant to the main thrust of the associated complaint (i.e., the document had nothing to do with the causes of action in the lawsuit or investigation). The question of scope arose in particular for production requests associated with the one complaint that on its face did not involve allegations against the tobacco industry (but which was instead about medical devices). Topic 49, which coupled that complaint with a production request for "[a]ll documents created between 1962 and 1999 referencing or including warnings or draft warnings used in the United States," proved to be particularly problematic because it was read by the assessor as being aimed at warnings for faulty medical devices. Not surprisingly, no relevant documents were found for topic 49. It was therefore removed from the evaluation because topics with no known relevant documents can not be used to compare the effectiveness of alternative system designs. Results are therefore reported for the remaining 39 topics.

As is often the case, assessors found some unintended ambiguity in the topics, either due to grammatical construction of the topic (e.g., what did the word "their" refer to), or due to inherent ambiguity embedded within words or concepts (e.g., what constitutes "lobbying efforts," "advertising," "marketing," and "promotion"). For one assessor, the word "event" (in a topic asking for all documents relating to the placement of product logos at events held in California), prompted them to consult the Random House Dictionary, where the word is defined as "something that occurs in a certain place during a particular interval of time." Therefore, in this assessor's view, documents that mentioned such activities as the America's Cup Race, speed skiing, auto racing, Hispanic Cultural events, Swing jam weekend, an antiviolence campaign, a country music festival, and an anti-smoking campaign called "Tobacco is Whacko," were all properly within the scope of the topic.

Another miscellaneous concern of one or more assessors involved how to deal with documents containing foreign language text. The track coordinators instructed assessors to make judgments based on English portions of documents, or otherwise mark the document as unsure.

In general, assessors took their jobs very seriously. A number of assessors made a second pass through their document set to resolve anomalies or to revisit judgments based on knowledge gained on the first pass. Many requests were directed to the track coordinators for help in resolving technical concerns.

It turned out that a nontrivial portion of the documents in the judgment pools could not be assessed at all using the assessment platform. While the same set of UCSF XML records provided the starting point for both the IIT CDIP version 1.0 collection and for the assessment platform's database, a few records with formatting problems were inadvertently treated differently by the two groups. In addition, a substantial number of XML records with variant formatting could not be loaded until assessment was already underway. More importantly, an even larger number of documents could not have their page images displayed during much of the assessment period. The total number of documents affected was less than 5% of the total collection, although somewhat more than 5% of the assessment pools were affected because longer documents were more likely to be affected. We addressed these problems by asking assessors to view documents at the LTDL Web site *(http://legacy.library.ucsf.edu/)* if their images could not be viewed on the CDIP platform, and record their assessments using the CDIP platform. In a very few cases, no record at all was loaded on the CDIP platform and assessments were sent by email. Also in a very few cases document images were found to be partial or missing on the LTDL Web site as well. In those few cases, assessors were asked to make a judgment based on the metadata record if possible, or to mark the document as "unsure."

The track coordinators asked assessors to record how much time they spent in performing assessment review. Based on post-assessment survey responses and related emails, assessment time data is available for 16 participants representing 39% of the overall assessment effort (12,743 of the 32,738 assessments). The reported review rate of documents reviewed per hour ranged from a low of 12.33 (Topic 31) to a high of 67.5 (Topic 25). The average review rate constituted 24.7 documents per hour. Note that each of the assigned topics included within it a highly varied set of documents, in terms of both differences in subject matter complexity as well as in total length.

4.3. Inter-Assessor Agreement

In order to assess the effects of differing assessor interpretations, we performed a limited amount of dual assessment after completion of the first phase of assessments. A sample of 50 documents (25 that had been judged as relevant, and 25 that had been judged as not relevant) was drawn from the pool for each of the 40 assessed topics. (Topic 49 was included for dual assessments, even though it could not be used for evaluating systems.) When fewer than 25 relevant documents had been identified, the number of non-relevant documents was increased to keep the total at 50. These sets were then assessed by a different assessor, without knowledge of the previous judgments. A total of 12 volunteers assessed documents in this second round, seven first-round veterans who received new topics to review, plus five new recruits.

Figure 1 shows the values of Cohen's kappa (Shoukri, 2004, Sec. 3.3), a chance-corrected measure of agreement, for each topic, as computed from the sample of 50 documents. Let:

n_{00} = number documents judged nonrelevant by main and secondary assessor,
n_{01} = number documents judged nonrelevant by main, but relevant by secondary,
n_{10} = number documents judged relevant by main, but nonrelevant by secondary, and
n_{11} = number of documents judged relevant by both main and secondary assessor.

where $n = n_{00} + n_{01} + n_{10} + n_{11}$ is for us equal to *50*. To compute kappa, one first computes the observed proportion of agreement between the assessors:

$$p_o = (n_{00} + n_{11}) / n$$

and the proportion agreement expected by chance under the assumption the assessors make their judgments independently with their particular observed frequencies of relevant and nonrelevant:

$$p_e = (n_{00} + n_{01})(n_{00} + n_{10}) / n^2 + (n_{10} + n_{11})(n_{01} + n_{11}) / n^2.$$

Cohen's kappa is then:

$$K = (p_o - p_e)/(1 - p_e).$$

The mean value of kappa over the 40 topics was +0.49, indicating moderate overall agreement between assessors (kappa ranges between -1 for complete disagreement to +1 for complete agreement), although considerable variation was evident across topics. The kappa values shown in Figure 1 are based on a sample of documents with (usually) 25 documents that the main assessor judged positive, and 25 they judged negative. The kappa value would have been different if a random sample from the pool had been judged by both assessors. We can compute an approximation of what kappa on the pool would have been by treating the 50 documents as a stratified sample and computing the expected values of the four contingency table cells that go into kappa. This is not quite an unbiased estimate of what kappa would have been on the pool, since kappa is a nonlinear function of the contingency table cells, but it is a reasonable approximation. Table 1 (which can be found at the end of this paper) shows the raw values of the contingency table entries along with kappa and other associated statistics. Table 2 (also at the end of the paper) shows the stratified estimates of what the contingency table cells would be for the full pool, along with approximations to the agreement measures computed by plugging the expected values of the contingency table cells into the formula for each measure.

As Voorhees has shown, moderate inter-annotator agreement can yield comparisons that are stable when one set of assessments are substituted for the other (Voorhees 2000). Evaluation measures should, therefore, be interpreted on a comparative rather than an absolute basis.

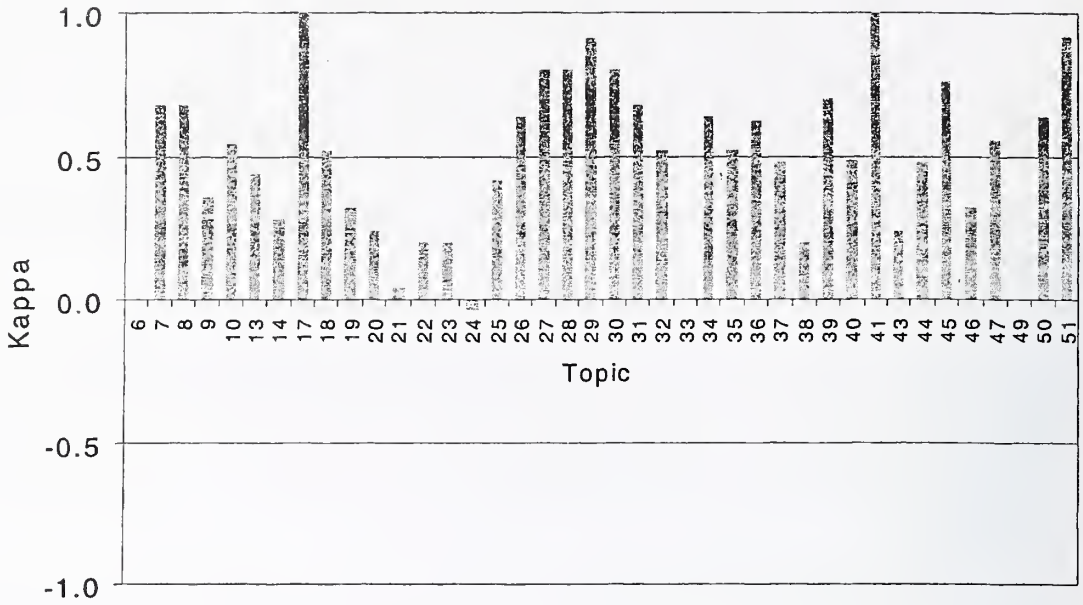


Figure 1. Chance-corrected inter-annotator agreement, by topic.

5. Results

Six participating sites submitted 31 ranked runs with no more than 5,000 documents per topic. Three of those runs applied a Boolean restriction when producing the document sets—those three runs consisted of substantially fewer than 5,000 documents for some topics. The baseline Boolean run, on the other hand, was not required to be ranked (although in practice it was first subjected to the Boolean constraint and then resulting Boolean set was ranked), so no upper bound on the size of the retrieved set was imposed in that case. The actual sizes of the submitted sets for the baseline Boolean run varied from 1 to 128,195 documents across topics. In addition to these 32 runs, the sets of approximately 100 documents found by the human expert for each topic (described in Section 4.1) were scored as if they were a 33rd run, (although as described below this comparison is not a fair one). Runs were given names beginning with an abbreviation that identified the submitting site. In this section, we briefly review the techniques used by each site; additional details can be found in the papers posted on the TREC Web site (http://trec.nist.gov).

- Hummingbird (hum). Hummingbird (now Open Text Corporation) submitted eight runs that explored the effects of alternative ways of formulating queries, different choices of index terms, and blind relevance feedback, plus the reference Boolean run (humL06B). The documents were indexed using the Livelink ECM-eDocs SearchServer system. The OCR field was indexed in every case, and all metadata was indexed together with OCR for seven runs, including the reference Boolean run (the exceptions being humL06dvo and humL06tvo). Queries were constructed automatically in six cases (the exceptions being humLo6B—the reference Boolean run, humL06t—the same run with a cutoff at 5,000, and humL06t0—a contrastive Boolean run using the first query in the negotiation history rather than the last query). For five of those six runs, the queries were automatically constructed from words in the Boolean queries (but without the use of Boolean or proximity operators); for the sixth run (humL06dvo) the queries were automatically constructed from the production request field.

- National University of Singapore (NUS). The National University of Singapore submitted two runs to explore the effects of evidence combination from multiple topic fields. The contents of the OCR field were indexed using the Lucene text retrieval system, and queries were formed from words found in the production request and the Boolean queries (but without the use of Boolean or proximity operators).

- Sabir Research (Sab). Sabir Research submitted seven runs to explore the effects of vocabulary filtering on OCR indexing and blind relevance feedback. The contents of the OCR and all metadata fields were indexed together using a vector space text retrieval system with pivoted document length normalization. Queries were formed from words in the production request and words in the Boolean Query for five of those runs; one run used only words from the production request (SabLeg06ar1) and one run used words from the production request, words from the Boolean query (without Boolean or proximity operators) and words from the Complaint (SabLeg06aa1).

- University of Maryland (Umd). The University of Maryland submitted four runs that explored the effects of different sources of query terms. The contents of the OCR and all metadata fields were indexed together using the Indri text retrieval system. Queries were formulated automatically for three runs: UmdBase (from words in the production request field), UmdBoolAuto (from words found in the final Boolean query, but without Boolean or proximity operators), and UmdComb (from both). For the fourth run (UmdBool), Indri queries were manually constructed to approximate the Boolean operators as closely as possible using Indri's query language (which does not directly support some required operators).

- University of Missouri-Kansas City (UMKC). The University of Missouri-Kansas City submitted eight runs that explored the effects of blind relevance feedback. The contents of the OCR field were indexed using the Lucene text retrieval system. Queries were formed automatically from words in the Boolean query (with Boolean operators, and sometimes also with proximity operators).

- York University (york). York University submitted two runs that explored the effects of blind relevance feedback. The contents of the OCR and all metadata fields were indexed together using Okapi BM 25 term weights. Queries were formulated automatically from words found in the Boolean query negotiation history (but without Boolean or proximity operators).

- Expert manual searcher "run" (EXPMANUAL). As described in section 4.1, the expert manual searcher used an interactive search system to identify up to 100 documents per topic that she felt would be unlikely to be retrieved by fully automated systems.

5.1. Uniques Analysis

One way of characterizing the results of different approaches to searching is to examine the contribution of each approach to the total set of known relevant documents. Figure 2 shows one way of looking at those statistics. As the grey bars show, on average across the 39 topics, 57% of the known relevant documents[3] were found by the reference Boolean query (i.e., either uniquely by the reference Boolean system, or by the reference Boolean system and also one or more other systems). As the analysis in Section 5.3 shows, our pooling strategy results in an underestimate of the actual number of relevant documents found by the reference Boolean system for topics with large numbers of relevant documents. Nevertheless, we this serves as a useful reference point from which to start an analysis of documents uniquely retrieved by other techniques.

The black bars stacked above the grey bars show the additional relevant documents found by the expert manual searcher but not by the reference Boolean system. On average across the 39 topics, the expert searcher found an additional 11% of the known relevant documents. In this case, the counts are accurate, since every document added to the pools by the expert searcher was judged. From this, we can conclude that by reformulating their query the expert searcher was able to find a substantial number of relevant documents that were not found by the reference Boolean system.

[3] In this section, and through the paper, the "known relevant documents" that we refer to are those judged as relevant by the primary assessor. Documents identified as relevant only by the second assessor in the inter-annotator agreement studies were not treated as relevant in the uniques analysis or when computing effectiveness metrics.

The white bars stacked above the black and grey bars show the additional relevant documents that were found by some system other then the reference Boolean system or the expert manual searcher. On average across the 39 topics, these other systems found an additional 32% of the known relevant documents. Our pooling strategy, which focuses on documents near the top of at least one ranked list and which includes no more than 100 documents from any one system, likely underestimates the number of relevant documents that ranked retrieval systems can find. Indeed, results for the "depth probe" run reported in the Hummingbird (Open Text) paper suggest that this underestimate may be substantial for at least some topics. Nonetheless, we can state with confidence that there were a large number of known relevant documents (1,417 across 39 topics) that were not found by the reference Boolean system or by the expert searcher. There was, therefore, scope for ranked retrieval systems to substantially outperform both the reference Boolean system and the expert manual searcher because there were a substantial number of known relevant documents that neither of those systems found. As we will see below, that did not happen.

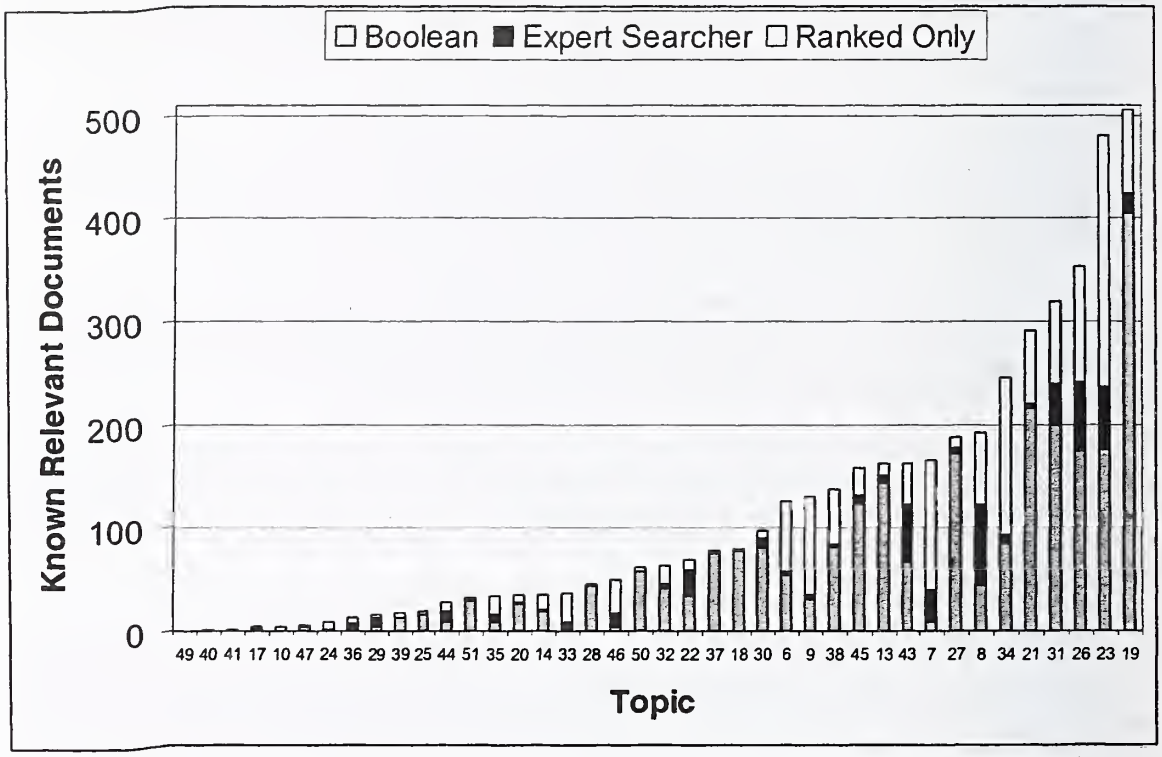


Figure 2. Known relevant documents found by the Reference Boolean system (grey), found by the expert searcher but not the reference Boolean system (black), and found uniquely by at least one other system (white).

5.2. R-Precision

Although our principal focus is on recall rather than precision, it is convenient to begin with a precision-oriented measures because precision-oriented measures are well understood, widely reported, and easily computed. Figure 3 compares the ranked retrieval runs using mean R-precision, a precision-oriented measure computed as the average across topics of the density of relevant documents at rank R (where R is the number of known relevant documents for that topic). The seven dark bars show the best scoring run from each participating team (and from the manual searcher). For comparison, all other runs (in order: the expert manual search, the reference Boolean run, and three Boolean runs from participating teams) are shown to the left of the ranked runs. Because R-precision is focused early in the ranked list, this measure would be expected to favor ranked retrieval systems. All four Boolean runs were, however, ranked in some way after being subjected to the Boolean constraint. The result is, therefore, in some sense fair in those cases. The expert human searcher "run" is disadvantaged in this comparison, however. It consisted of only

about 100 documents, those documents were not intentionally ranked by probability of relevance, and the searcher focused on finding diverse relevant documents to enrich the pool rather than the easiest relevant documents to boost measured effectiveness.

Three results are clearly evident in this data. First, the best runs from three of the participating sites were nearly indistinguishable by the R-precision measure, and one of those three runs (humL06t) was subjected to a Boolean constraint. Indeed, the reference Boolean run did about as well on this precision-oriented measure as the best unconstrained ranked retrieval runs. This is notable because Boolean runs can retrieve only documents that satisfy the Boolean query, while the ranked runs had no such constraint. From this we can conclude that (when averaged over 39 topics), little adverse effect resulted from respecting the Boolean constraint. Of course, with only six participating systems we are nowhere near exhausting the design space for search techniques, so ways may yet be found to achieve improvements that are not available to a Boolean system. All we can say at this point is that such improvements have not yet been demonstrated in the TREC legal track.

The second obvious result is that Boolean systems are not all created equal—two of the four Boolean runs did about twice as well (by this precision-oriented measure) as the other two! In one case (Hummingbird) this appears to result from using the initial rather than the final Boolean queries. In the other case (Maryland) the differences appear to result from incomplete support for extended Boolean operators. When we first proposed this track, one of our shorthand goals was to see if someone could "beat Boolean." This year's results indicate that might be easily achieved in the wrong way (by inadvertently creating an underperforming "Boolean" baseline), and that careful attention to the process by which the Boolean queries are created and used will be important if we are to produce meaningful comparisons.

Third, the expert manual searcher's submitted sets had, despite the factors discussed above that would tend to decrease R-precision scores, noticeably higher R-precision than any of actual submitted runs (all of which were essentially fully automatic, although in a few cases some query reformatting was done manually). This suggests that focusing attention on interactive search might yield interesting results.

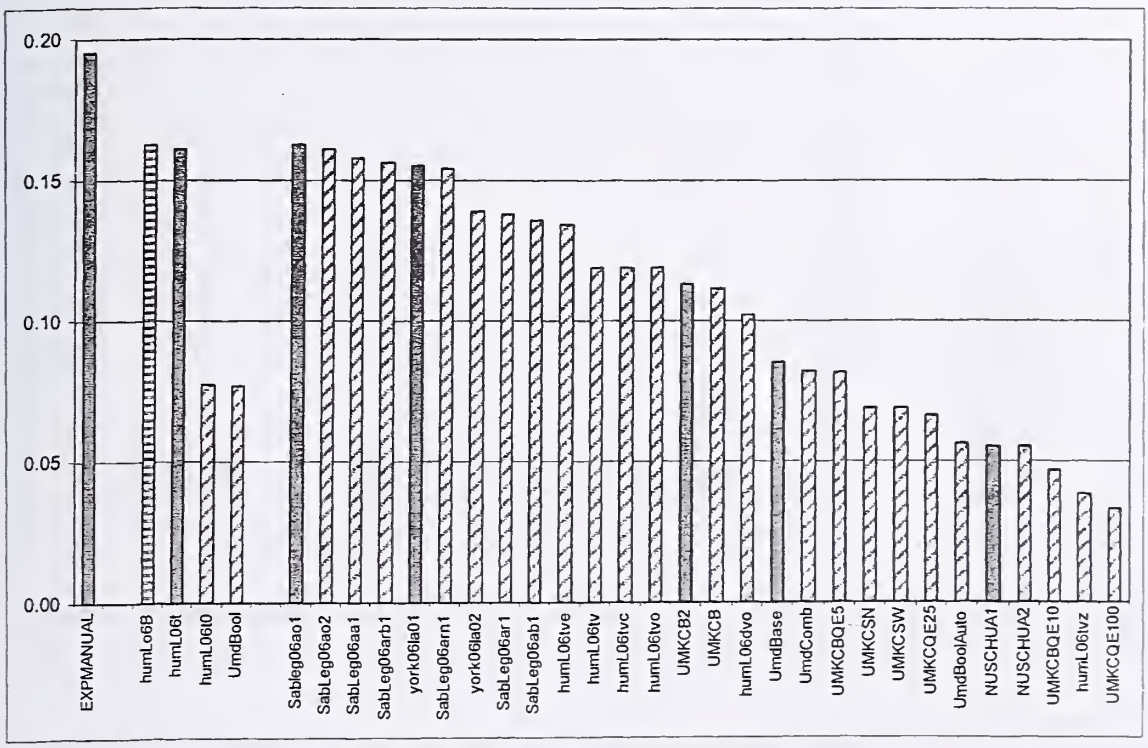


Figure 3. Mean precision at R (the actual number of known relevant documents for each topic). Ranked runs on left side, Reference runs on right side. Best run for each team shown as solid bar.

Runs EXPMANUAL and humL06tvz were not conventionally ranked and thus are disadvantaged by this measure.

5.3. P@B

A set-oriented comparison of ranked retrieval with the reference Boolean run was possible for 22 topics for which 5,000 or fewer documents were included in the Boolean set.[4] Let B be the size of the submitted set for the baseline Boolean run for a particular topic. The idea is to treat the top B documents of a ranked run for that topic as if it were a submitted set of size B and then compute P@B, the density of relevant documents in that set (treating unassessed documents as not relevant). Although the true number of relevant documents is not known, the precision at any fixed cutoff is proportional to the recall at that same cutoff, so we can interpret P@B for any individual topic as a measure of recall. Averaging across topics yields somewhat different results than a direct computation of recall would, however, since the constant of proportionality varies by topic.

In Figure 4 we compare P@B values for SabLeg06ao2 (one of the top-scoring runs by P@R) with those of the baseline Boolean run. For 12 of 22 topics, P@B favors the reference Boolean run, while for 7 of 22 the ranked run is favored. Three topics had tied values of P@B that were near 0.

The above analysis understates the true value of P@B since the assessed pools are incomplete and biased in favor of documents ranked highly by submitted runs. This problem is worse for a set-based measure like P@B than for measures like R-precision that focus on the documents closest to the top of a ranked list. We had no alternative to pooling for evaluating the ranked run, but for the baseline Boolean run an unbiased estimate of P@B could be computed using stratified sampling.

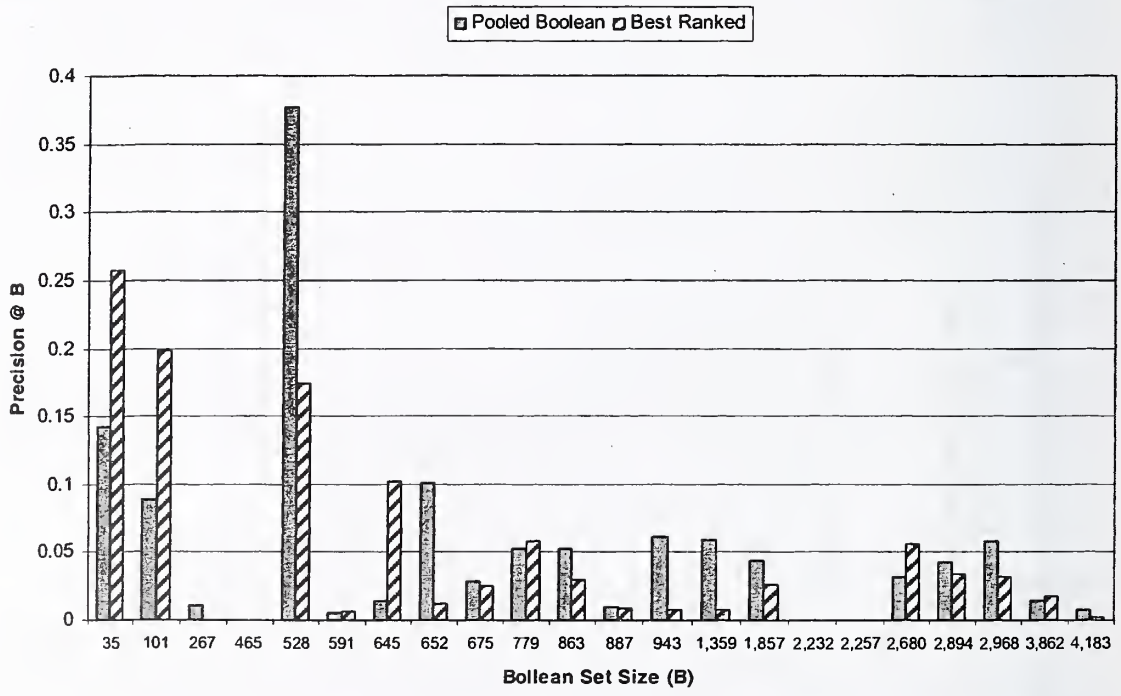


Figure 4. Recall-oriented effectiveness measure, by topic, in increasing order of Boolean set size. Topic 33 (for which B=1) not shown.

[4] There were actually 23 topics with B≤5,000, but using topic 33, for which B=1, would not be informative because when B=1 precision can only be 0 or 1. Precision at B for topic 33 was 0 for the reference Boolean run, and 1 for the best ranked run.

It turned out that the identity of the original stratified samples (Section 4.1) from the baseline Boolean run could not be recovered at the time that evaluation measures were computed because of a hardware failure. Further, the original stratification could not be reconstructed from the pools themselves because documents meeting the strata definitions could have come from ranked runs, the expert manual run, or the stratified sampling process. However, we were able to define new strata in a way that still allowed the computation of unbiased estimates of set-based effectiveness measures for the baseline Boolean run.

We separated the documents in the baseline Boolean set for each topic into four strata based on which other runs they occurred in:

- ***Stratum 0'***: Documents occurring in the top 100 of any site's main run, top 10 of any run from any site, or in the expert manual set.
- ***Stratum 1'***: Documents in former ***Stratum 1***, but not in ***Stratum 0'***.
- ***Stratum 2'***: Documents in former ***Stratum 2***, but not in ***Stratum 0'***.
- ***Stratum 3'***: Documents in former ***Stratum 3***, but not in ***Stratum 0'***.

By putting all documents added to the pool by a run other than the baseline Boolean run into ***Stratum 0'***, we can treat any remaining documents as if they had been drawn randomly from the newly defined strata. ***Stratum 0'*** is treated as having all its documents sampled, while the number of documents treated as sampled from ***Strata 1', 2',*** and ***3'*** varies by topic. We used the resulting stratified samples to produce unbiased estimates of P@B for the baseline Boolean run, as well as computing a 95% confidence interval for these estimates using the Gaussian approximation to the binomial (Lewis, 1996). Because these new strata generally contain fewer documents than under the original stratification, our estimates of P@B usually have a higher sampling variance than they would have with the original stratification.

As Figure 5 shows, pooling and stratified sampling produce the same estimate of P@B when B is at or below 267. The situation is quite different as B grows, however. In 10 of the 16 cases for which B is 528 or higher, and for which the pooled estimate of P@B is nonzero, the pooled estimate falls below the lower limit of the confidence interval on the stratified estimate. This result reinforces our earlier that our pool-based effectiveness measures do not provide a measure of the absolute effectiveness of any of the participating systems. Further, the large gap between the pool-based P@B and the true value (or at least an unbiased estimate of it) means more danger that biases in pool construction will affect even comparisons of relative effectiveness.

Analysis reported in the Hummingbird (Open Text) paper indicates that similar effects are present in at least the one ranked "depth probe" run for which a kind of stratified sampling was done (humL06tvz). Our future work on comparison of ranked and Boolean runs will require a more nuanced strategy than we have yet applied.

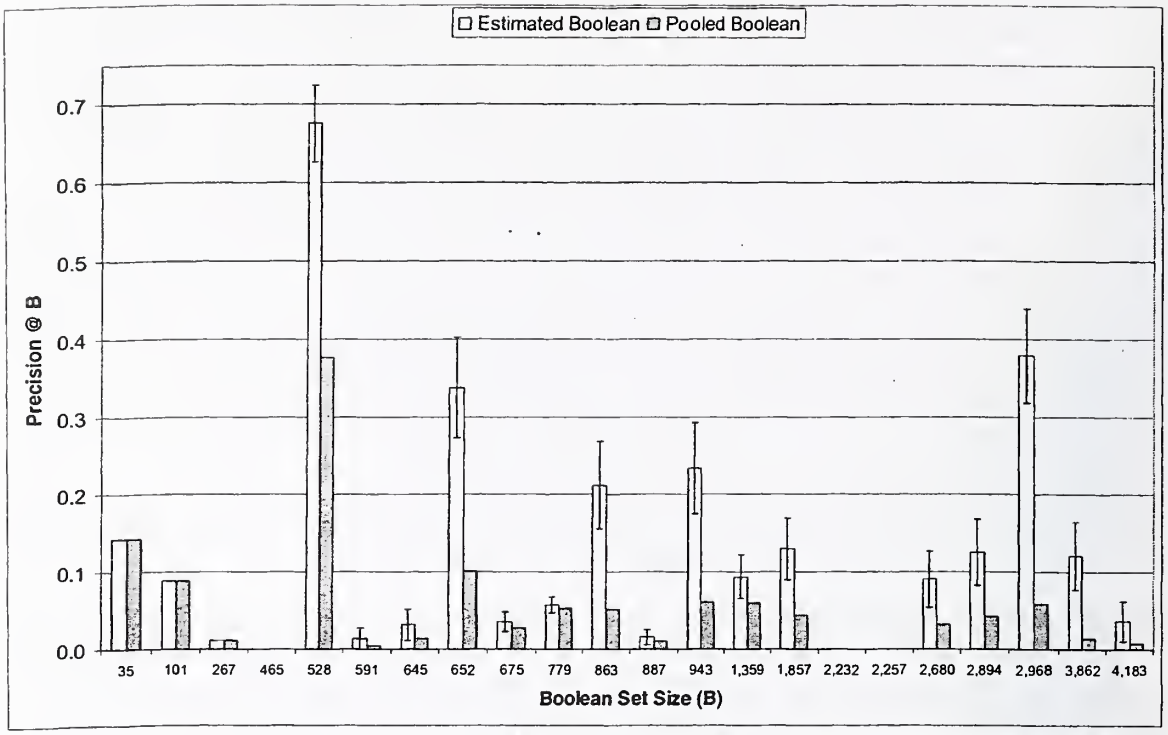


Figure 5. Comparison of stratified estimate of P@B with pool-based estimate of P@B.

6. Conclusion

This first year of the TREC Legal Track has produced a new test collection that models present practice in e-discovery, and that will also be of interest to researchers working on retrieval from scanned document images and to researchers working on the integrated use of structured metadata with document text as a basis for retrieval. Six research teams participated in the evaluation, contributing to the creation of relevance assessment pools that were judged in a manner representative of the human review process that precedes release in an e-discovery process. These judgments provide a basis for both this year's evaluation and for development of new approaches that are tuned to the unique characteristics of this task.

Analysis of the results yielded a number of useful insights. Perhaps the most striking result is the strong performance of the Boolean queries. The reference Boolean run did about as well (by R-precision) as the best ranked runs, and the top seven ranked runs (again, by R-precision) all used terms from the Boolean queries as part of their automatic query formulation process. This suggests that the negotiated Boolean queries are information-rich, which has implications both for practice (propounding Boolean queries is a productive activity) and for system design (leveraging manually constructed Boolean queries when they are available can yield improved retrieval effectiveness). A second important result is objectively quantifying the fact that there are many relevant documents to be found beyond those identified by strict application of negotiated Boolean queries. This should not be surprising, of course, since it is well known that formulating queries that are both sufficiently inclusive and sufficiently precise is difficult. Perhaps the most important implication of this observation is that exploring system designs based on relaxation of the Boolean query and based on augmenting queries using terms from other sources (e.g., the production request) may ultimately yield better retrieval effectiveness than strict application of Boolean logic. While that potential was not realized in the TREC 2006 legal track (at least not by the P@R measure), this year's relevance judgments are exactly what is needed to explore the space of possible system designs to determine whether such gains can indeed be achieved.

From the perspective of evaluation design, the clearest conclusion is that additional work on statistical estimation for recall-oriented measures is needed. The analyses in this paper and in the Open Text paper

indicate that statistical estimates of retrieval effectiveness for both the reference Boolean run and for one ranked run yield markedly different results from the more commonly used metrics in which unassessed documents are treated as not relevant. Additional analysis will be needed before we can directly compare those two runs, and the potential for statistical estimation for other ranked retrieval runs from 2006 is limited by the sampling strategies that were employed when forming the assessment pools. It will therefore be important to revisit both our choice of measures and our sampling strategies for the 2007 Legal Track.

Our focus in this first year of the Legal Track was on the design of automated systems, but of course automated systems are ultimately used by people. Our expert searcher run yielded some interesting insights, however, finding an average of 13 documents per topic that the reference Boolean query had missed and achieving better retrieval effectiveness (by the P@R measure) than any other run. This suggests that a focused effort to explore interactive search techniques in the TREC 2007 legal track might yield additional insights.

Perhaps the greatest accomplishment of the TREC 2006 Legal Track is that it happened at all. More than 50 volunteers contributed to assembling and distributing the collection, creating topics, developing systems, managing submissions, creating pools, judging relevance, developing metrics, creating scoring software, analyzing results, and coordinating those activities. This has yielded the results that we would hope for from any TREC track in its first year: (1) a reusable test collection to support future research, (2) a set of baseline results to which future research can be compared, and (3) a community of researchers who bring a variety of perspectives to these important challenges. The coordinators trust that a second year of research will continue to yield important results.

Acknowledgements

This track would not have been possible without the generous support of the IIT CDIP project (including Gady Agam, Shlomo Argamon, Ophir Frieder, and Dave Grossman, plus special thanks to David Roberts), the University of California at San Francisco Library (particularly Karen Butter, Albert Jew, Kirsten Neilsen, and Heidi Schmidt), members of the Sedona Conference®, and the volunteer relevance assessors and their participating firms and institutions. In particular, the coordinators wish to thank Ryan Bilbrey, Conor Crowley, Joe Looby, and Stephanie Mendelsohn, for their greatly appreciated assistance in writing draft complaints, in topic development, and for participating in "Boolean negotiations," and Anna Marshall at George Washington University School of Law for her extraordinary assessor recruitment efforts. Special acknowledgement is due Richard Braman, Executive Director of The Sedona Conference ®, for all of his assistance in facilitating a successful outcome of year 1 of the Legal Track. Thanks also to Michael Tacelosky and Keith Ivey of Tobacco Documents Online and *smokefree.net* for access to and help with their collection of tobacco documents and for their work on our assessment platform. Finally, special thanks also go to our colleagues at NIST for handling much of the logistics.

References

Agam, G., Argamon, S., Frieder, O., Grossman, D. and Lewis, D., "Complex Document Information Processing: Prototype, Test Collection, and Evaluation," *Document Recognition and Retrieval XIII,* SPIE Proceedings vol. 6067, pp. 60670N-1 to 60670N-11, 2006.

Baron, J., "Toward a Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery," *Sedona Conference Journal*, vol. 6, pp. 237-246, 2005 (available from Westlaw and LEXIS)

Baron, J., Lewis, D., and Oard, D. " "How To" Guide for Assessors – TREC Legal Track 2006." Version 4, Aug. 20, 2006. *http://trec-legal.umiacs.umd.edu/TRECLegal_HowToGuide_Version4Final.doc*

Blair, D., and Maron, M. "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System," *Communications of the ACM*, 28(3)289-299, 1985.

Buckley, C., Dimmick, D., Soboroff, I and Voorhees, E., "Bias and the Limits of Pooling," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 619-620, Seattle, 2006

Buckley, C. and Voorhees, E. "Retrieval System Evaluation," in *TREC: Experiment and Evaluation in Information Retrieval*, E. M. Voorhees and D. K. Harman, eds., MIT Press, pp. 53-75, 2002.

Cochran, W. *Sampling Techniques*, 3rd edition. John Wiley & Sons, New York, 1977.

Coleman v. Morgan Stanley, 2005 WL 679071 (Fla. Cir. Ct. Mar. 1, 2005)

Collaborative Expedition Workshop #45, *Advancing Information Sharing, Access, Discovery and Assimilation of Diverse Digital Collections Governed by Heterogeneous Sensitivities*, held Nov. 8, 2005, see http://colab.cim3.net/cgi-bin/wiki.pl?AdvancingInformationSharing_DiverseDigitalCollections_HeterogeneousSensitivities_11_08_05

Jacobs, P., "GE in TREC-2: Results of a Boolean Approximation Method for Routing and Retrieval," in *The Second Text Retrieval Conference (TREC-2)*, Gaithersburg, MD, August, 1993, pp. 191-200.

Lewis, D. "The TREC-5 Filtering Track," in *The Fifth Text Retrieval Conference (TREC-5)*, Gaithersburg, MD, November, 1996, pp. 75-96.

Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D., and Heard, J. "Building a Test Collection for Complex Document Information Processing," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 665-666, Seattle, 2006.

Lu, X., Allan, J., Miller, D. "Boolean Systems Revisited: Its Performance and Behavior," *The Fourth Text Retrieval Conference (TREC-4)*, pp. 459-474, Gaithersburg, MD, November, 1995.

Schmidt, H.; Butter, K.; and Rider, C. "Building Digital Tobacco Document Libraries at the University of California, San Francisco Library/Center for Knowledge Management," *D-Lib Magazine*, 8(2), 2002.

Sedona Conference, *The Sedona Principles: Best Practices Recommendations & Principles for Addressing Electronic Document Production* (2005 version), Principle 11, see http://www.thesedonaconference.org/content/miscFiles/publications_html

Turtle, H., "Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance," *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 212-220, Dublin, 1994.

U.S. Federal Rules of Civil Procedure, Rules 26 & 34, as amended (Dec. 1, 2006)

Voorhees, E., "Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness," *Information Processing and Management*, 36(5)697-716, 2000.

Zubulake v. UBS Warburg, 217 F.R.D. 309 (S.D.N.Y. 2004)

Topic	n	n11	n01	n10	n00	Agree	Agree R	Agree N	Kappa
6	50	0	0	25	25	0.5	0	0.667	0
7	50	21	4	4	21	0.84	0.84	0.84	0.68
8	50	19	2	6	23	0.84	0.826	0.852	0.68
9	49	9	0	16	24	0.673	0.529	0.75	0.355
10	50	2	0	3	45	0.94	0.571	0.968	0.545
13	50	11	0	14	25	0.72	0.611	0.781	0.44
14	50	11	4	14	21	0.64	0.55	0.7	0.28
17	50	4	0	0	46	1	1	1	1
18	50	20	7	5	18	0.76	0.769	0.75	0.52
19	50	8	0	17	25	0.66	0.485	0.746	0.32
20	50	7	1	18	24	0.62	0.424	0.716	0.24
21	50	5	4	20	21	0.52	0.294	0.636	0.04
22	50	5	0	20	25	0.6	0.333	0.714	0.2
23	50	9	4	16	21	0.6	0.474	0.677	0.2
24	50	0	1	9	40	0.8	0	0.889	-0.037
25	50	7	6	5	32	0.78	0.56	0.853	0.414
26	50	21	5	4	20	0.82	0.824	0.816	0.64
27	50	22	2	3	23	0.9	0.898	0.902	0.8
28	50	21	1	4	24	0.9	0.894	0.906	0.8
29	50	16	1	1	32	0.96	0.941	0.97	0.911
30	50	22	2	3	23	0.9	0.898	0.902	0.8
31	50	23	6	2	19	0.84	0.852	0.826	0.68
32	50	20	7	5	18	0.76	0.769	0.75	0.52
33	50	0	0	25	25	0.5	0	0.667	0
34	50	20	4	5	21	0.82	0.816	0.824	0.64
35	50	14	1	11	24	0.76	0.7	0.8	0.52
36	50	9	3	4	34	0.86	0.72	0.907	0.627
37	50	14	2	11	23	0.74	0.683	0.78	0.48
38	50	17	12	8	13	0.6	0.63	0.565	0.2
39	50	15	4	3	28	0.86	0.811	0.889	0.7
40	50	1	2	0	47	0.96	0.5	0.979	0.485
41	50	1	0	0	49	1	1	1	1
43	50	10	4	15	21	0.62	0.513	0.689	0.24
44	50	12	0	13	25	0.74	0.649	0.794	0.48
45	50	19	0	6	25	0.88	0.864	0.893	0.76
46	50	8	0	17	25	0.66	0.485	0.746	0.32
47	50	4	3	2	41	0.9	0.615	0.943	0.558
49	50	0	32	0	18	0.36	0	0.529	0
50	50	19	3	6	22	0.82	0.809	0.83	0.64
51	50	24	1	1	24	0.96	0.96	0.96	0.92
MEAN						0.765	0.627	0.810	0.490

Table 1: Raw contingency table entries from interassessor comparison study. We show agreement, i.e. *(n00 + n11)/n*, agreement on relevant, i.e. *2*n11 / (2*n11 + n01 + n10)*, agreement on nonrelevant, i.e. *2*n00 / (2*n00 + n01 + n10)*, and Cohen's kappa.

Top	pool	E[n11]	E[n01]	E[n10]	E[n00]	~E[Agree]	~E[AgreeR]	~E[AgreeN]	~E[Kappa]
6	840	0	0	125	715	0.851	0	0.92	0
7	854	138.6	110.2	26.4	578.8	0.84	0.67	0.894	0.57
8	857	145.9	53.2	46.1	611.8	0.884	0.746	0.925	0.671
9	849	46.8	0	83.2	719	0.902	0.529	0.945	0.488
10	858	2	0	3	853	0.997	0.571	0.998	0.57
13	837	71.3	0	90.7	675	0.892	0.611	0.937	0.559
14	716	15.8	108.8	20.2	571.2	0.82	0.197	0.899	0.129
17	767	4	0	0	763	1	1	1	1
18	769	64	192.9	16	496.1	0.728	0.38	0.826	0.263
19	919	161.6	0	343.4	414	0.626	0.485	0.707	0.298
20	938	9.8	36.1	25.2	866.9	0.935	0.242	0.966	0.209
21	893	58.2	96.3	232.8	505.7	0.631	0.261	0.754	0.046
22	853	13.8	0	55.2	784	0.935	0.333	0.966	0.315
23	832	173.2	56.3	307.8	295.7	0.563	0.487	0.619	0.183
24	924	0	22.3	9	892.7	0.966	0	0.983	-0.014
25	961	11.1	148.7	7.9	793.3	0.837	0.124	0.91	0.092
26	935	297.4	116.2	56.6	464.8	0.815	0.775	0.843	0.62
27	916	165.4	58.2	22.6	669.8	0.912	0.804	0.943	0.747
28	910	38.6	34.6	7.4	829.4	0.954	0.648	0.975	0.625
29	875	16	26	1	833	0.969	0.542	0.984	0.529
30	781	85.4	54.7	11.6	629.3	0.915	0.72	0.95	0.672
31	707	294.4	92.9	25.6	294.1	0.832	0.832	0.832	0.668
32	770	51.2	197.7	12.8	508.3	0.727	0.327	0.828	0.225
33	570	0	0	37	533	0.935	0	0.966	0
34	810	196	90.4	49	474.6	0.828	0.738	0.872	0.611
35	542	19	20.3	15	487.7	0.935	0.519	0.965	0.484
36	872	9	69.7	4	790.3	0.916	0.196	0.955	0.175
37	863	43.7	62.8	34.3	722.2	0.887	0.474	0.937	0.412
38	741	93.2	289.9	43.8	314.1	0.55	0.358	0.653	0.118
39	887	15	108.6	3	760.4	0.874	0.212	0.932	0.183
40	832	1	33.9	0	797.1	0.959	0.056	0.979	0.053
41	876	1	0	0	875	1	1	1	1
43	820	64.8	105.3	97.2	552.7	0.753	0.39	0.845	0.236
44	821	13.4	0	14.6	793	0.982	0.649	0.991	0.641
45	755	120.1	0	37.9	597	0.95	0.864	0.969	0.834
46	627	16	0	34	577	0.946	0.485	0.971	0.464
47	733	4	49.6	2	677.4	0.93	0.134	0.963	0.121
49	983	0	629.1	0	353.9	0.36	0	0.529	0
50	756	47.1	83.3	14.9	610.7	0.87	0.49	0.926	0.426
51	936	31.7	36.2	1.3	867.8	0.96	0.628	0.979	0.61
mean	825					0.854	0.462	0.901	0.396

Table 2: Stratified estimates of what the interassessor agreement contingency table values would be on the full pools, along with approximate expected values of agreement, agreement on relevant, agreement on nonrelevant, and kappa.

Overview of the TREC 2006 Question Answering Track

Hoa Trang Dang[1], Jimmy Lin[2], and Diane Kelly[3]

[1]National Institute of Standards and Technology
Gaithersburg, MD 20899
hoa.dang@nist.gov

[2]University of Maryland
College Park, MD 20742
jimmylin@umd.edu

[3]University of North Carolina
Chapel Hill, NC 27599
dianek@email.unc.edu

Abstract

The TREC 2006 question answering (QA) track contained two tasks: the main task and the complex, interactive question answering (ciQA) task. As in 2005, the main task consisted of series of factoid, list, and "Other" questions organized around a set of targets; in contrast to previous years, the evaluation of factoid and list responses distinguished between answers that were globally correct (with respect to the document collection), and those that were only locally correct (with respect to the supporting document). The ciQA task provided a framework for participants to investigate interaction in the context of complex information needs, and was a blend of the TREC 2005 QA relationship task and the TREC 2005 HARD track. Multiple assessors were used to judge the importance of information nuggets used to evaluate the responses to ciQA and "Other" questions, resulting in an evaluation that is more stable and discriminative than one that uses only a single assessor to judge nugget importance.

1 Introduction

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 natural language question. Since its inception in TREC-8 (1999), the track has steadily expanded both the type and difficulty of the questions asked. 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 contained list and definition questions in addition to factoid questions (Voorhees, 2004). A list question asks for different answer instances that satisfy the information need, such as *List the names of chewing gums.* Answering such questions requires a system to assemble a response 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.

In TREC 2004 (Voorhees, 2005a), factoid and list questions were grouped into different series, where each series was associated with a target (a person, organization, or thing) and the questions in the 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 was roughly equivalent to the definition questions in the TREC 2003 task.

Since the beginning of the QA track, the document returned with an answer had been used to determine the time frame for a question. For example, "Ronald Reagan" was considered a correct answer for the question *Who is the President of the United States?* if that answer was supported by a document from 1987, even if more recent documents supported "George Bush" as the answer. Such guidelines were appropriate because questions were primarily phrased in the present tense without specifying an explicit time frame. However, in the TREC 2005 main task, events were added as a possible target for the question series, and it became clear that the time frame implied by the series could not be ignored when judging the correctness of answers. Event targets and temporally-constrained questions required that questions be interpreted in the temporal context explicit in the question or implicit in the series.

The main task for the TREC 2006 QA track was the same as the main task in 2005, except that the implicit time frame for questions phrased in the present tense was the date of the last document in the document collection, rather than the document returned with the answer. Thus, systems were required to give the most up-to-date answer supported by the document collection. This restriction brought TREC QA more closely in line with question answering in the real world, where users would want the best answer to their question in the document collection, rather than just any answer found in any document. The evaluation of the question series in 2006 also down-weighted factoid questions, which had been tested for many years, by giving equal weight to each of the 3 question types in the final per-series score.

In addition to the main task, the TREC 2006 QA track also contained a complex, interactive QA (ciQA) task. The 2006 ciQA task was a blend of the TREC 2005 relationship task (Voorhees and Dang, 2006) and the TREC 2005 HARD track, which focused on single-iteration clarification dialogues (Allan, 2006). The goals of the ciQA task were to push the frontiers of question answering away from "factoid" questions towards more complex information needs that exist within richer user contexts, and to move away from the one-shot interaction model implicit in previous evaluations towards a model based at least in part on interactions with users. Two metrics were introduced to evaluate answers to complex questions in the ciQA task: modified F-scores based on nugget pyramids and recall plots based on response length.

The remainder of this paper describes each of the two tasks in the TREC 2006 QA track in more detail. Section 2 describes the questions, evaluation methods, and results for the main task, while Section 3 discusses the ciQA task. The final section looks at the future of the track.

2 Main Task

The scenario for the main task in the TREC 2006 QA track was that an adult, native speaker of English was looking for information about a target of interest. The target could be a person, organization, thing, or event. The user was assumed to be an "average" reader of U.S. newspapers. Serving as surrogate users, NIST assessors developed the questions and judged the system responses.

The main task required systems to provide answers to a series of related questions. A question series, which focused on a target, consisted of several factoid questions, one to two list questions, and exactly one Other question. The order of questions 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 final test set contained 75 series; the targets of these series are given in Table 1. Of the 75 targets, 19 were PERSONs, 19 were ORGANIZATIONs, 19 were EVENTs, and 18 were THINGs. The series contained a total of 403 factoid questions, 89 list questions, and 75 Other questions. Each series contained 6–9 questions (counting the Other question), with most series containing 8 questions.

Participants were required to submit 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 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 the same series, but could not "look ahead" and use later questions to help answer earlier questions. Thus, question series can be viewed as an abstraction of an information-seeking dialogue between the user and the system; cf. (Kato et al., 2004). The document collection from which answers were to be drawn was the AQUAINT Corpus of English News Text (LDC catalog number LDC2002T31). As a convenience for track participants, NIST made available document rankings of the top 1000 documents per target as produced using the PRISE document retrieval system, with the target as the query. In total, 59 runs from 27 participants were

145	John William King convicted of murder		
	145.1	FACTOID	How many non-white members of the jury were there?
	145.2	FACTOID	Who was the foreman for the jury?
	145.3	FACTOID	Where was the trial held?
	145.4	FACTOID	When was King convicted?
	145.5	FACTOID	Who was the victim of the murder?
	145.6	LIST	What defense and prosecution attorneys participated in the trial?
	145.7	OTHER	
185	Iditarod Race		
	185.1	FACTOID	In what city does the Iditarod start?
	185.2	FACTOID	In what city does the Iditarod end?
	185.3	FACTOID	In what month is it held?
	185.4	FACTOID	Who is the founder of the Iditarod?
	185.5	LIST	Name people who have won the Iditarod.
	185.6	FACTOID	How many miles long is the Iditarod?
	185.7	FACTOID	What is the record time in which the Iditarod was won?
	185.8	LIST	Which companies have sponsored the Iditarod?
	185.9	OTHER	
212	Barry Manilow		
	212.1	FACTOID	What year was he born?
	212.2	FACTOID	How many times has he married?
	212.3	FACTOID	What is the name of the musical that he wrote about the Harmonistas?
	212.4	FACTOID	What music school did he attend?
	212.5	FACTOID	For what female singer was he the musical director and pianist in the 70's?
	212.6	FACTOID	What record label did he sing for in 2000?
	212.7	LIST	List the songs he recorded.
	212.8	OTHER	

Figure 1: Sample question series from the test set. Series 145 has an EVENT as the target, series 185 has a THING as the target, and series 212 has a PERSON as the target.

141	Warren Moon	179	Hedy Lamarr
142	LPGA	180	Lebanese Parliament
143	American Enterprise Institute	181	Manchester United Football Club
144	82nd Airborne Division	182	1998 Edinburgh Fringe
145	John William King convicted of murder	183	Thabo Mbeki elected president of South Africa
146	Pakistani government overthrown in 1999	184	1999 Chicago Marathon
147	Britain's Prince Edward marries	185	Iditarod Race
148	tourists massacred at Luxor in 1997	186	Pyramids of Egypt
149	The Daily Show	187	Amazon River
150	television show Cheers	188	avocados
151	Winston Cup	189	Joanne Kathleen Rowling
152	Wolfgang Amadeus Mozart	190	H. J. Heinz Co.
153	Alfred Hitchcock	191	International Rowing Federation
154	Christopher Reeve	192	Basque ETA
155	Hugo Chavez	193	World Food Program (WFP)
156	NASCAR	194	1996 World Chess Super Tournament
157	United Nations (U.N.)	195	East Timor Independence
158	Tufts University	196	Adoption of the Euro
159	Wal-Mart	197	cloning of mammals (from adult cells)
160	IMF	198	Bushehr Nuclear Facility
161	1999 Baseball All-Star Game	199	Padre Pio
162	Multiple Myeloma	200	Frank Sinatra
163	Hermitage Museum	201	William Shakespeare
164	Judi Dench	202	Cole Porter
165	the Queen Mum's 100th Birthday	203	Nissan Corp.
166	avian flu outbreak in Hong Kong	204	Church of Jesus Christ of Latter-day Saints (Mormons)
167	the Millennium Wheel	205	1991 eruption of Mount Pinatubo
168	Prince Charles' paintings	206	Johnstown flood
169	stone circles	207	Leaning Tower of Pisa
170	John Prine	208	Great Wall of China
171	Stephen Wynn	209	Carolyn Bessette Kennedy
172	Ben & Jerry's	210	Janet Reno
173	World Tourism Organization (WTO)	211	Patsy Cline
174	American Farm Bureau Federation (AFBF)	212	Barry Manilow
175	repatriation of Elian Gonzales	213	Meg Ryan
176	An Officer and a Gentleman	214	2000 Miss America Pageant
177	Deep Blue	215	1999 Sundance Film Festival
178	methamphetamine labs		

Table 1: Targets of the 75 question series.

submitted to the main task.

The evaluation of a single run can be decomposed into component evaluations for each of the question types and a final per-series score. Each of the three question types has its own response format and evaluation method. The individual component evaluations in 2006 were identical to those used in the TREC 2005 QA track, except that a distinction was made between locally correct answers (supported in the associated document, but contradicted in later documents in the collection) and globally correct answers. An aggregate score was computed for each series in a run using a simple average of the component scores of questions in that series, and the final score for the run was computed as the average of its per-series scores.

2.1 Factoid questions

The system response to 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 made the final determination. Each response was assigned exactly one of the following five judgments:

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

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

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

locally correct: the answer string consists of exactly a correct answer that is supported by the document returned, but a more recent document contradicts the answer;

globally correct: the answer string consists of exactly the correct answer, that answer is supported by the document returned, and there are no later documents that contradict the answer.

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 if the question asks about "the Taj Mahal"). Questions also had to be interpreted in the time frame implied by the question series. For example, if the target was the event "France wins World Cup in soccer" and the question was *Who was the coach of the French team?* then the correct answer must be "Aime Jacquet", the name of the coach of the French team in 1998 when France won the World Cup, and not just the name of any past or current coach of the French team. NIL responses were correct only if there was no known answer to the question in the collection. NIL was correct for 17 of the 403 factoid questions in the test set. For 26 questions, no system returned the correct answer, although those questions did have a correct answer found by the assessors.

The main evaluation metric for the factoid component was *accuracy*, the fraction of questions judged to be globally correct. Table 2 shows the most accurate run for the factoid component for each of the top 10 groups. 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; NIL recall is the ratio of the number of times NIL was returned and correct to the number of times it was correct in the entire test set (17). If NIL was never returned, NIL precision is undefined and NIL recall is zero.

2.2 List questions

A list question asks for different instances of a particular type. The correct answer for a list question is the set of all such distinct instances in the document collection. A system's response to a list question consists of an unordered set of [*doc-id*, *answer-string*] pairs such that each *answer-string* represents a correct answer instance.

During the evaluation process, the assessor was given an entire system's run at a time. Each instance was evaluated in the same manner as the factoid questions, i.e., assigned one of the following judgments: incorrect, unsupported, not exact, locally correct, and globally correct. In addition to judging for correctness, the assessor also marked the answer instances for distinctness. The assessor arbitrarily chose any one of equivalent responses to be distinct, and the remainder were considered not distinct. Thus, systems were not rewarded (and in fact, penalized) for returning equivalent answer instances multiple times. Only globally correct responses could be marked as distinct.

The final set of globally correct answers for a list question was compiled from the union of distinct globally correct answers across all runs plus instances the assessor found during question development. For the 89 list questions in the test set, the average number of answers per question was 10, with a minimum of 2 and a maximum of 50. A system's response to a list question was scored using instance precision (IP) and instance recall (IR) based on the complete list of known distinct instances. Let S be the number of such instances, D be the number of globally 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 to produce an F-score 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 89 questions. Table 3 gives the average F-score of the run with the best list component score for each of the top 10 groups.

2.3 Other questions

The Other questions were evaluated using the methodology originally developed for the TREC 2003 definition questions. A system's response for an Other question consisted of an unordered set of [*doc-id*, *answer-string*] pairs. The answer strings were presumed to contain interesting "nuggets" about 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 more difficult than answering TREC 2003 definition questions.

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

In the second step of the evaluation process, the assessor went through each system's output in turn and marked which nuggets appeared in the response. An answer string contained a nugget if there was a *conceptual* match between the answer string and the nugget; that is, the match was independent of the particular wording used in either the nugget or the system output. A nugget match was marked at most once per response—if the system output 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, nugget recall was computed as the ratio of the number of matched nuggets to the total number of vital nuggets in the list. Nugget precision was much more difficult to compute since there was no effective way of enumerating all the concepts contained in a particular answer string. Instead, a measure based on length (in non-whitespace characters) was used as an approximation to nugget precision. The length-based measure granted an allowance of 100 characters for each (vital or non-vital) nugget matched. If the total system output was less than this number of characters, the value of nugget precision was 1.0. Otherwise, the measure's value decreased as the length increased according to the following formula:

$$1 - \frac{length - allowance}{length}.$$

The final score for an Other question was an F-score, with nugget recall weighted more heavily than nugget precision:

$$F(\beta) = \frac{(\beta^2 + 1) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}.$$

The score for the Other questions component was the average F-score (β=3) over the 75 Other questions. Table 4 gives the F-score for the best scoring Other question component for each of the top 10 groups.

2.3.1 Nugget Pyramids

The vital/okay distinction has previously been identified as a weakness in the TREC nugget-based evaluation methodology (Hildebrandt et al., 2004). Since only vital nuggets affect nugget recall, it is difficult for systems to achieve non-zero scores on topics with few vital nuggets in the answer key. Thus, scores are easily affected by assessor errors and other random variations in evaluation conditions. One direct consequence is that in previous TREC evaluations, the median score for many questions turned out to be zero (Voorhees, 2005b). A binary distinction on nugget importance is insufficient to discriminate between the quality of runs that return no vital nuggets but different numbers of okay nuggets. To address many of these issues, Lin and Demner-Fushman (2006) proposed an extension called "nugget pyramids", in which multiple assessors provide judgments of whether a nugget is vital or simply okay.

To examine the effectiveness of the pyramid approach, NIST also computed F-scores for Other responses using the pyramid extension. Nine different sets of vital/okay judgments were solicited from eight unique assessors (the primary assessor who originally created the nuggets later assigned vital/okay labels again). Each assessor was given all the questions for the series, as well as the nuggets created by the primary assessor. Using the pyramid procedure, a weight was assigned to each nugget based on the number of assessors who marked it as vital. These nugget weights were then incorporated into the nugget recall computation.

The left graph in Figure 2 plots the average F-scores for each run as computed using a single assessor vs. using the nugget pyramid. Even though the nugget pyramid does not represent any single real user, average pyramid F-scores do correlate highly with average single-assessor F-scores; the Pearson's correlation is 0.987, with a 95% confidence interval of [0.980, 1.00].

While the average F-score for a particular run is stable given a large enough number of questions, the F-score for a single Other question does vary depending on the assessor. The right graph in Figure 2 plots the single-assessor and pyramid F-scores for each individual Other question from all submitted runs. The Pearson correlation between single-assessor and pyramid F-scores in this case is 0.870, with a 95% confidence interval of [0.863, 1.00]. For 16.4% of the questions, the nugget pyramid assigned a non-zero F-score where the original single-assessor F-score was zero. Thus, from the perspective of system developers, the F-scores from the nugget pyramids may be more useful since they are more discriminative. For a more detailed analysis of the nugget pyramids extension, please refer to (Dang and Lin, 2007).

2.4 Per-series Combined Weighted Scores

The three component scores measure a system's ability to process each type of question, but may not reflect the system's overall usefulness to a user. Since each individual series is an abstraction of a single user's interaction with the system, taking the individual series as the basic unit of evaluation should provide a more accurate representation of the effectiveness of the system from an individual user's perspective. Since each series is a mixture of different question types, we can compute a weighted average of the scores of the three question types on a per-series basis, and take the average of the per-series weighted scores as the final score for the run (Voorhees, 2005b). In 2006, the weighted average of the three component scores for an individual series was computed as:

$$\text{WeightedScore} = \frac{1}{3} \times \text{Factoid} + \frac{1}{3} \times \text{List} + \frac{1}{3} \times \text{Other}.$$

To compute the weighted score for an individual series, only the scores for questions belonging to that series were included in the computation. Since each of the component scores ranges between 0 and 1, the weighted score is also in that range. In contrast to previous years, when factoid questions were weighted more heavily than the other questions,

Run Tag	Submitter	Accuracy	NIL Prec	NIL Recall
lccPA06	Language Computer Corporation (Moldovan)	0.578	0.000	0.000
LCCFerret	Language Computer Corporation (Harabagiu)	0.538	–	0.000
cuhkqaepisto	The Chinese University of Hong Kong	0.390	0.107	0.353
ed06qar1	University of Edinburgh	0.323	0.069	0.294
InsunQA06	Harbin Institute of Technology (HIT)	0.298	0.118	0.353
QACTIS06A	National Security Agency (NSA)	0.266	0.118	0.118
ILQUA1	University of Albany	0.266	0.027	0.059
NUSCHUAQA1	National University of Singapore	0.261	0.000	0.000
asked06c	Tokyo Institute of Technology	0.251	–	0.000
QASCU3	Concordia University (Kosseim)	0.213	0.000	0.000

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

Run Tag	Submitter	F
lccPA06	Language Computer Corporation (Moldovan)	0.433
cuhkqaepisto	The Chinese University of Hong Kong	0.188
NUSCHUAQA1	National University of Singapore	0.171
FDUQAT15A	Fudan University (Wu)	0.165
QACTIS06C	National Security Agency (NSA)	0.156
LCCFerret	Language Computer Corporation (Harabagiu)	0.148
ILQUA1	University of Albany	0.129
Roma2006run3	University of Rome "La Sapienza"	0.127
csail02	Massachusetts Institute of Technology (MIT)	0.125
InsunQA06	Harbin Institute of Technology (HIT)	0.118

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

Run Tag	Submitter	F($\beta = 3$)
ed06qar1	University of Edinburgh	0.250
FDUQAT15A	Fudan University (Wu)	0.223
QASCU3	Concordia University (Kosseim)	0.199
lccPA06	Language Computer Corporation (Moldovan)	0.167
uw574	University of Washington (UW CLMA group)	0.164
Roma2006run3	University of Rome "La Sapienza"	0.164
MITRE2006C	The MITRE Corp.	0.156
QACTIS06C	National Security Agency (NSA)	0.154
NUSCHUAQA3	National University of Singapore	0.150
ISL2	University of Karlsruhe & Carnegie Mellon University	0.150

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

equal weight was given to the three components in 2006. The final per-series score of each run is simply the average of individual per-series scores.

Table 5 shows the final per-series score for the best run from each group. We fit a two-way analysis of variance model with the target type and the best run from each group as factors, and the final per-series score as the dependent variable; we found significant differences between target types ($p = 0.005$) and runs (p essentially equal to 0). To determine which runs were significantly different from each other, we performed a multiple comparison using Tukey's honestly significant difference criterion and controlling for the experiment-wise Type I error so that the probability of declaring a difference between two runs to be significant when it is actually not, is at most 5%. Table 5 shows the results of the multiple comparison; runs sharing a common letter are not significantly different. A similar multiple comparison showed that PERSON targets had significantly higher scores than EVENTs, but no significant differences between any of the other target types were found.

System scores on the main task have declined since TREC 2004 even though the question series format of the main task has been the same. This is not surprising given that the questions have become increasingly more difficult, with "simple" factoid questions requiring higher levels of reasoning to extract the correct answer from the documents. Assessors also have become more strict about disallowing inexact answers as correct answers.

3 The Complex, Interactive QA (ciQA) Task

The goal of the complex, interactive question answering (ciQA) task is to push the frontiers of question answering in two directions:

- A move away from "factoid" questions towards more complex information needs that exist within richer user contexts. (Question series in the main task also exemplify this shift in evaluation focus.)
- A move away from the one-shot interaction model implicit in previous evaluations towards a model based at least in part on interactions with users.

In terms of implementation, the 2006 ciQA task was a blend of the TREC 2005 relationship task (Voorhees and Dang, 2006) and the TREC 2005 HARD track, which focused on single-iteration clarification dialogues (Allan, 2006).

3.1 Complex "Relationship" Questions

The complex information needs explored by ciQA represented an extension and refinement of so-called "relationship" questions piloted in TREC 2005. This choice provided some continuity and training data for participants.

The concept of a "relationship" is defined as the ability of one entity to influence another, including both the means to influence and the motivation for doing so. Evidence for both the existence or absence of ties is relevant. The particular relationships of interest naturally depend on the context.

A relationship question in the ciQA task, which we refer to as a topic, is composed of two parts. Consider an example:

> **Template:** What evidence is there for transport of [drugs] from [Mexico] to [the U.S.]?
> **Narrative:** The analyst would like to know of efforts to curtail the transport of drugs from Mexico to the U.S. Specifically, the analyst would like to know of the success of the efforts by local or international authorities.

The question template is a stylized information need that has a fixed structure and free slots (items in square brackets) whose instantiation varies across different topics. The narrative is free-form natural language text that elaborates on the information need, providing, for example, user context, a more articulated statement of interest, focus on particular topical aspects, etc. Five template types were developed for the ciQA task, enumerated in Figure 3. For the final test set, NIST assessors developed a total of 30 topics, with 6 topics for each of these templates.

Answers to ciQA topics consisted of [*doc-id*, *answer-string*] pairs, and were evaluated using the same nugget-based methodology that was employed for the main task Other questions. However, the total length of system responses was limited to 7,000 non-whitespace characters. Two metrics were employed to quantify answer quality:

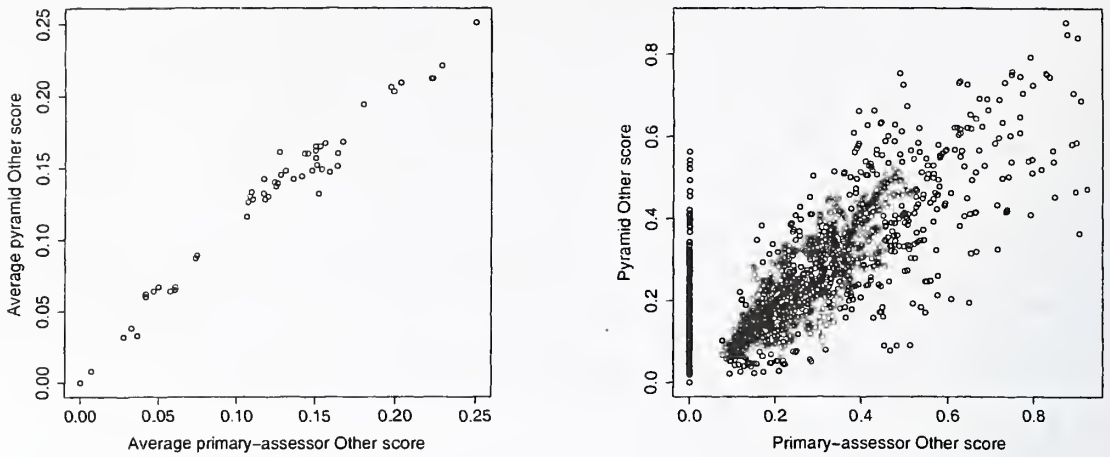


Figure 2: Other F-score computed using a single primary assessor vs. using multiple assessors, by individual question (right), and averaged over all questions for each submitted run (left).

RunID	Per-series score												
lccPA06	0.3938	A											
LCCFerret	0.2644		B										
cuhkqaepisto	0.2310		B	C									
ed06qar1	0.2066		B	C	D								
FDUQAT15A	0.1918			C	D	E							
NUSCHUAQA3	0.1908			C	D	E							
QACTIS06A	0.1853			C	D	E	F						
ILQUA1	0.1713				D	E	F	G					
QASCU1	0.1588				D	E	F	G	H				
Roma2006run3	0.1571				D	E	F	G	H				
InsunQA06	0.1568				D	E	F	G	H				
MITRE2006C	0.1485					E	F	G	H				
ISL2	0.1430					E	F	G	H	I			
csail02	0.1344					E	F	G	H	I			
shef06ss	0.1344					E	F	G	H	I			
lsv2006c	0.1298						F	G	H	I	J		
asked06c	0.1156							G	H	I	J		
uw574	0.1083								H	I	J	K	
DLT06QA02	0.0871									I	J	K	L
TIQA200601	0.0851									I	J	K	L
clr06m	0.0763										J	K	L
TWQA0601	0.0725										J	K	L
irstqa06	0.0573											K	L
Dal06e	0.0459												L
lexiclone06	0.0458												L
lf10w10g5	0.0312												L
TREC06ST01	0.0167												

M
M
M
M
M
M
M
M
M

Table 5: Multiple comparison of the best run from each group, based on ANOVA of per-series score.

What evidence is there for transport of [goods] from [entity] to [entity]?
Example: What evidence is there for transport of [drugs] from [Mexico] to [the U.S.]?

What [relationship] exist between [entity] and [entity]?
(where [relationship] ∈ {"financial relationships", "organizational ties", "familial ties", "common interests"})
Example: What [financial relationships] exist between [drug companies] and [universities]?

What influence/effect do(es) [entity] have on/in [entity]?
Example: What effect does [aspirin] have on [coronary heart disease]?

What is the position of [entity] with respect to [issue]?
Example: What is the position of [John McCain] with respect to [the Moral Majority or the Christian Coalition]?

Is there evidence to support the involvement of [entity] in [event/entity]?
Example: Is there evidence to support the involvement of [China] in [human organ transplants from Chinese prisoners]?

Figure 3: The five templates used in the TREC 2006 ciQA task.

- The first and primary metric was the F-score ($\beta = 3$) with the "nugget pyramid" extension.
- The second metric was new for the ciQA task and attempted to graphically capture the tradeoffs between conciseness and completeness (Lin, 2007). The basic idea is to quantify weighted nugget recall (what we call pyramid recall) as a function of answer length (in non-whitespace characters). By the nugget pyramid building process, each nugget is assigned a weight between zero and one. Weighted nugget recall is the sum of weights of all nuggets retrieved divided by the sum of all weights of all nuggets in the assessor's answer key.

Implementing this metric required two important changes to the previous evaluation protocol:

1. Answer strings must be rank ordered, with best first.
2. Assessors must mark the first instance of a nugget in the response set of answer strings.

For the recall plots, the scoring methodology was as follows (character counts do not include whitespaces):

1. For each topic, NIST recorded the cumulative character length and pyramid recall after each answer string had been assessed.
2. Each data point was interpolated to the nearest 100 character increment (longer than the current length). For example, a pyramid recall of 0.25 at 168 characters would be interpolated to (200, 0.25). Plotting these points yielded pyramid recall as a function of answer length for a particular topic.
3. To arrive at the recall plot for a particular system run, the mean of the recall values was taken across all topics at each length increment, i.e., mean pyramid recall over all topics at 100 characters, at 200 characters, at 300 characters, etc.

3.2 Interactive Question Answering

The purpose of the interactive aspect of ciQA was to provide a framework for participants to investigate interaction in the QA context and to provide an opportunity for non-QA researchers to become involved in this area. We consider an interactive system to be a system that gives users control over all or a portion of displayed content. Using this definition, the smallest possible interaction unit consists of the user responding to the system and the system using the

user's response to produce new content. The interactive aspect of ciQA was concerned with this interaction unit and was modeled in part after the HARD track's clarification forms.

The HARD track's clarification forms allowed participants to elicit information from assessors through a single interaction. This interaction consisted of assessors completing forms (i.e., Web pages) that had been created by track participants. The results of these interactions were then returned to the participants so that revised results could be generated—comparison of output before and after the clarification quantified the effects of the interaction.

Although many participants took advantage of the opportunity provided by the HARD track to investigate traditional relevance feedback techniques, this was not a goal of the HARD track nor a condition for participation; there were, in fact, some participants who used clarification forms in novel ways. In the ciQA task, we explicitly encouraged innovative ways of using forms that go beyond traditional relevance feedback. The question answering community has yet to reach common ground on the role of interaction in QA, and the ciQA task was meant to provide a forum for continued dialogue.

The rationale for studying the smallest interaction unit rests on the idea that a good QA system should return relevant information with a minimum amount of interaction. Furthermore, given the potential complexities that are likely to arise with coordinating cross-site interactive evaluations, we believe that using the smallest interaction unit is a reasonable starting point in the exploration of interactive QA. Previous experiences with the TREC interactive track demonstrated that coordinating multi-site interactive IR system evaluation is a challenge and that results are difficult (if not impossible) to compare.

In more detail, interaction forms were HTML pages created by participants that solicited user input via CGI. Although NIST placed no restrictions on the type of content, there were technical restrictions (see below). Each question was associated with a unique form, and each site was limited to two sets of interaction forms (which provided the ability to evaluate two different interaction techniques).

NIST assessors completed the interaction forms on Redhat Enterprise Linux workstations with 20-inch LCD monitors (1600×1200 resolution and millions of colors) using the Firefox Web browser (v1.5.0.2). The machines at NIST were disconnected from all networks and participants were required to provide all necessary information as part of their forms. If a form required multiple files, then it was necessary for such files to be contained within the submitted directory structure. These forms were not allowed to invoke any CGI scripts or write to disk. Javascript was allowed, but Java was not.

Assessors spent no more than three minutes completing each interaction form. This duration included the time needed to load the form, initialize any content, and then render it. At the end of three minutes, if the assessor had not submitted the form, the form timed out and was forcibly submitted. The CGI variable bindings associated with the forms captured the results of the interactions, which NIST returned to the participants.

3.3 Results

The ciQA evaluation proceeded as follows:

1. Participants submitted initial runs and interaction forms.
2. NIST assessors interacted with the forms.
3. NIST returned results of the interaction (i.e., the CGI bindings).
4. Participants submitted final runs based on the results of the interactions.
5. NIST evaluateed both initial and final runs.

As with the main task, the AQUAINT collection of newswire articles served as the official corpus. To support the individual goals of participants, ciQA was entirely independent of the main task; the interactive aspect was also optional, which allowed participants to focus solely on complex QA if they desired. Finally, both automatic and manual runs were allowed. A manual run was defined as any run where human intervention occurred in any part of the process (except assessor interaction with the submitted interaction forms).

The ciQA task drew participation from six groups. NIST received ten initial runs and eleven final runs. A total of ten sets of interaction forms were submitted by the six participants. In addition, a pair of initial/final runs that

used simple sentence retrieval techniques was submitted as a baseline implementation (described below). A set of interaction forms was also associated with this run pair.

We constructed a rotation specifying the order in which interaction forms would be presented to assessors to minimize learning and order effects, and to insure that each form would occupy each position in the rotation (e.g., first, second, third) as equal a number of times as possible. This rotation is shown in Table 6. Row headings show topic numbers, while column headings represent forms. Cell numbers indicate the presentation order of the form; for example, for Topic 26, CLR1 was the fourth form presented and strath3 was the first. This rotation is based on a basic Latin square rotation; the relationship between forms is preserved, but the position of the form is shifted across topics. For example, strath2 always followed CLR1 except when it was the first form in the rotation, and strath2 and CLR1 were each the first, second, third, etc. form in the rotation an equal number of times. To construct the order, forms and topics were randomly assigned to column and row headings, and an order of 1, 2, 3, 4, etc. was assigned to the first row, 2, 3, 4, 5, etc. was assigned to the second row, etc. The table has been sorted according to topic, but one can see that Topics 31, 34 and 39 appeared in either the 1st, 12th, or 23rd row of the randomized table.

In total, there were eleven different initial–final run pairs. The pyramid F-scores of these run pairs are shown in Table 7. Pyramid F-scores were computed using the methodology outlined in (Lin and Demner-Fushman, 2006). Nine different sets of vital/okay judgments were solicited from eight unique assessors (the assessor who originally created the nuggets later assigned vital/okay labels again).

In addition to runs submitted by the participants, the University of Maryland separately prepared a sentence retrieval baseline. For each topic, the verbatim question template was used as a query to Lucene, which returned the top 20 documents. These documents were then tokenized into individual sentences. Sentences that contained at least one non-stopword from the question were retained and returned as the initial run (up to the 7,000 character limit). Sentence order within each document and across the ranked list was preserved. The interaction forms associated with this run asked the assessor for relevance judgments on each of the sentences (relevant, not relevant, don't know). The final run was prepared by removing sentences judged not relevant—this had the effect of pulling in more sentences from documents lower in the ranked list. The performance of this sentence retrieval baseline is also shown in Table 7.

Surprisingly, the sentence retrieval baseline performed exceedingly well. Only two initial runs received a higher score, one of which was a manual run. Only two final runs received a higher score, one of which was a manual run. The high baseline performance is consistent with findings from previous TREC results (Voorhees, 2004). Figure 4 shows a scatter plot of the initial and final F-scores for all eleven run pairs. Points below the reference line $y = x$ represent cases in which interaction actually decreased performance—there were two such cases.

Plots of pyramid recall as a function of response length are shown in Figure 5. These graphs attempt to quantify how quickly a user is able to acquire relevant nuggets by reading system responses. Naturally, curves that rise more quickly represent "better" systems. In the top graph, the sentence retrieval baseline is compared against the best automatic run. In the bottom graph, the sentence retrieval baseline is compared against the best manual run. It is interesting to note that for the automatic runs, these recall plots paint a different picture of performance than the pyramid F-scores. Although UWATCIQA4 achieved a higher pyramid F-score than the final submission of the sentence retrieval baseline, the recall plots suggest that the sentence retrieval baseline is able to deliver more information given the same response length. For the manual run, although the recall plots show little difference between the nugget content of the pre- and post-interaction system responses, the pyramid F-scores suggest a difference in answer quality. More work is needed to understand the divergences between pyramid F-scores and these recall plots.

These results appear to suggest that the complex QA task is difficult, but that off–the–shelf IR systems provide a strong baseline. The effective use of linguistic analysis techniques for complex questions remains an open research question. For a more in-depth exploration of these issues and the evaluation methodology, see (Lin, 2007).

4 Future of the QA Track

At the TREC 2006 workshop, participants indicated that they would like to have longer, more complex interactions in the ciQA task rather than short interactions via cached interaction forms. Participants proposed trying "live interactions" for 2007. Under this setup, the interactive QA system would be located at a URL on the participant's machine, and NIST assessors would simply navigate to the URL. The advantage would be that participants would be able to host more complex interaction interfaces. On the other hand, this setup would put additional burden on each partici-

Topic	CLR1	strath2	csaili2	UMDA1	UMAS1	UWAT1	CLR2	csaili1	strath3	UMDM1	Baseline1
26	4	5	6	7	8	9	10	11	1	2	3
27	7	8	9	10	11	1	2	3	4	5	6
28	4	5	6	7	8	9	10	11	1	2	3
29	9	10	11	1	2	3	4	5	6	7	8
30	9	10	11	1	2	3	4	5	6	7	8
31	1	2	3	4	5	6	7	8	9	10	11
32	6	7	8	9	10	11	1	2	3	4	5
33	3	4	5	6	7	8	9	10	11	1	2
34	1	2	3	4	5	6	7	8	9	10	11
35	5	6	7	8	9	10	11	1	2	3	4
36	5	6	7	8	9	10	11	1	2	3	4
37	6	7	8	9	10	11	1	2	3	4	5
38	5	6	7	8	9	10	11	1	2	3	4
39	1	2	3	4	5	6	7	8	9	10	11
40	11	1	2	3	4	5	6	7	8	9	10
41	2	3	4	5	6	7	8	9	10	11	1
42	10	11	1	2	3	4	5	6	7	8	9
43	8	9	10	11	1	2	3	4	5	6	7
44	10	11	1	2	3	4	5	6	7	8	9
45	11	1	2	3	4	5	6	7	8	9	10
46	8	9	10	11	1	2	3	4	5	6	7
47	2	3	4	5	6	7	8	9	10	11	1
48	8	9	10	11	1	2	3	4	5	6	7
49	7	8	9	10	11	1	2	3	4	5	6
50	11	1	2	3	4	5	6	7	8	9	10
51	6	7	8	9	10	11	1	2	3	4	5
52	3	4	5	6	7	8	9	10	11	1	2
53	10	11	1	2	3	4	5	6	7	8	9
54	7	8	9	10	11	1	2	3	4	5	6
55	9	10	11	1	2	3	4	5	6	7	8

Table 6: Form rotation according to topic. As a specific example: for Topic 26, the form CLR1 was presented fourth and the form strath3 was presented first.

Organization	Type	Run tags		Pyramid F-Score	
		Initial	Final	Initial	Final
CL Research	automatic	clr06ci1	clr06ci1r	0.151	0.184
CL Research	automatic	clr06ci2	clr06ci2r	0.175	0.209
MIT	automatic	csail1	csailif1	0.203	0.209
MIT	automatic	csail1	csailif2	0.203	0.203
U. Maryland	automatic	UMDA1pre	UMDA1post	0.224	0.180
U. Maryland	manual	UMDM1pre	UMDM1post	0.316	0.350
U. Mass.	automatic	UMASSauto2	UMASSi2	0.171	0.160
U. Mass.	automatic	UMASSauto1	UMASSi1	0.133	0.150
U. Strathclyde	manual	strath1	strath4	0.227	0.239
U. Waterloo	automatic	UWATCIQA1	UWATCIQA3	0.247	0.247
U. Waterloo	automatic	UWATCIQA1	UWATCIQA4	0.247	0.268
Baseline	automatic	-	-	0.237	0.264

Table 7: Performance of the eleven initial–final pairings for the ciQA task, along with the sentence retrieval baseline.

pant; if the NIST assessor could not reach a site for any reason during the interaction period – even due to problems outside the control of the site – the assessor would simply ignore the site. A straw poll indicated preference for live interactions, and the ciQA task will be repeated in 2007 with live URLs and a longer interaction period. Based on the successful application of the nugget pyramid evaluation method in TREC 2006, the pyramid method will be the official evaluation method for both the ciQa and the Other questions in TREC 2007.

Since the main task had been run largely unchanged for three years, a radical change was proposed to push the state of the art forward. The series format has supported the evaluation of different types of questions (factoid, list and Other) while providing an abstraction of a real user session with a QA system; therefore, rather than changing the series format, it was decided to move the main task forward by changing the genre of the document collection. The main task for the TREC 2007 QA Track will again be series of factoid, list, and Other questions, but the document collection will be a combination of newswire and blogs. Mining blogs for answers will introduce significant new challenges in at least two aspects that are very important for functional QA systems: 1) being able to handle language that is not well-formed, and 2) dealing with discourse structures that are more informal and less reliable than newswire.

References

James Allan. 2006. HARD track overview in TREC 2005: High accuracy retrieval from documents. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*.

Hoa Trang Dang and Jimmy Lin. 2007. Different structures for evaluating answers to complex questions: Pyramids won't topple, and neither will human assessors. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*.

Wesley Hildebrandt, Boris Katz, and Jimmy Lin. 2004. Answering definition questions with multiple knowledge sources. In *Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004)*, pages 49–56.

Tsuneaki Kato, Jun'ichi Fukumoto, Fumito Masui, and Noriko Kando. 2004. 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.

Jimmy Lin and Dina Demner-Fushman. 2006. Will pyramids built of nuggets topple over? In *Proceedings of the 2006 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2006)*, pages 383–390.

Jimmy Lin. 2007. Is question answering better than information retrieval? A task-based evaluation framework for question series. In *Proceedings of the 2007 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2007)*, pages 212–219.

Ellen M. Voorhees and Hoa T. Dang. 2006. Overview of the TREC 2005 question answering track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*.

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

Ellen M. Voorhees. 2005a. Overview of the TREC 2004 question answering track. In *Proceedings of the Thirteenth Text REtreival Conference (TREC 2004)*, pages 52–62.

Ellen M. Voorhees. 2005b. Using question series to evaluate question answering system effectiveness. In *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 299–306.

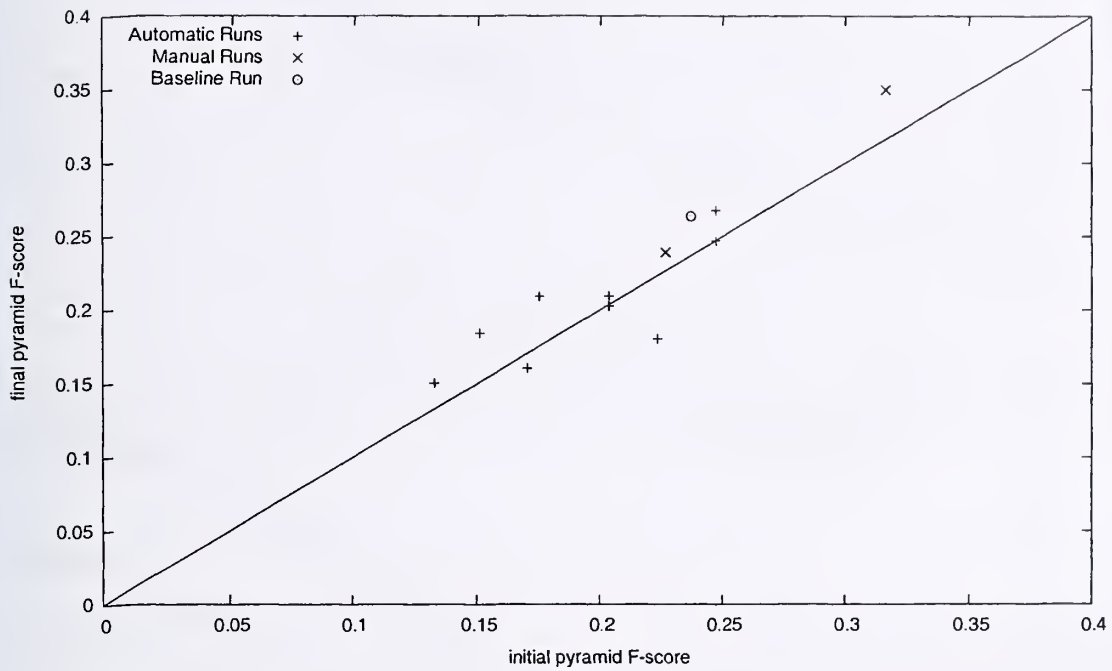


Figure 4: Scatter plot showing initial and final pyramid F-scores for submitted ciQA runs.

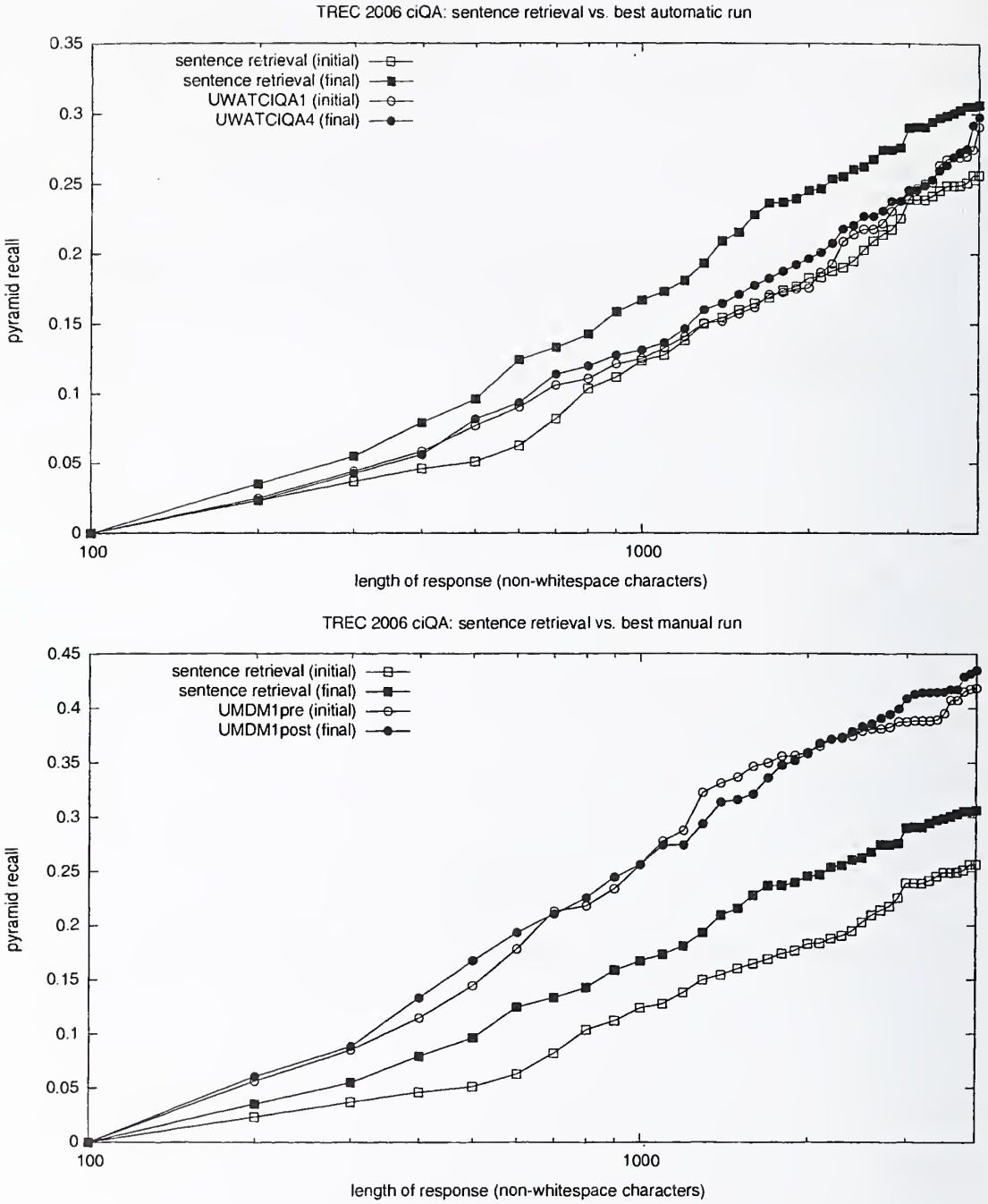


Figure 5: Plots of pyramid recall from ciQA runs as a function of response length: sentence retrieval baseline vs. the best automatic run (top) and vs. the best manual run (bottom)

TREC 2006 Spam Track Overview

Gordon Cormack
University of Waterloo
Waterloo, Ontario, Canada

1 Introduction

TREC's *Spam Track* uses a standard testing framework that presents a set of chronologically ordered email messages a spam filter for classification. In the filtering task, the messages are presented one at at time to the filter, which yields a binary judgement (*spam* or *ham* [i.e. non-spam]) which is compared to a human-adjudicated *gold standard*. The filter also yields a *spamminess* score, intended to reflect the likelihood that the classified message is spam, which is the subject of post-hoc ROC (Receiver Operating Characteristic) analysis. Two forms of user feedback are modeled: with *immediate feedback* the gold standard for each message is communicated to the filter immediately following classification; with *delayed feedback* the gold standard is communicated to the filter sometime later, so as to model a user reading email from time to time in batches. A new task – *active* learning – presents the filter with a large collection of unadjudicated messages, and has the filter request adjudication for a subset of them before classifying a set of future messages. Four test corpora – email messages plus gold standard judgements – were used to evaluate subject filters. Two of the corpora (the *public* corpora, one English and one Chinese) were distributed to participants, who ran their filters on the corpora using a track-supplied toolkit implementing the framework. Two of the corpora (the *private* corpora) were not distributed to participants; rather, participants submitted filter implementations that were run, using the toolkit, on the private data. Nine groups participated in the track, each submitting up to four filters for evaluation in each of the three tasks (filtering with immediate feedback; filtering with delayed feedback; active learning).

2 Spam Track Tasks

Broadly speaking, there were two spam track tasks: filtering, in which participant filters performed on-line classification with simulated user feedback; and active learning, in which participant filters were given a large number of historical email messages and allowed to request adjudication by the user for some of them before classifying a set of new messages. Task guidelines and tools may be found on the web at http://plg.uwaterloo.ca/~gvcormac/spam/ .

2.1 Filtering – Immediate Feedback

The immediate feedback filtering task is identical to the TREC 2005 task[1]. A chronological sequence of messages is presented to the filter using a standard interface. The the filter classifies each message in turn as either *spam* or *ham,* also computes a *spamminess score* indicating its confidence that the message is spam. The test setup simulates an ideal user who communicates the correct (gold standard) classification to the filter for each message immediately after the filter classifies it.

Participants were supplied with tools, sample filters, and sample corpora (including the TREC 2005 public corpus) for training and development. Filters were evaluated on four new corpora developed for TREC 2006.

2.2 Filtering – Delayed Feedback

Real user's don't immediately report the correct classification to filters. They read their email some time, typically in batches, some time after it is classified. The delayed filtering task simulates this delay in the following manner: the filter is asked to classify some number of messages without feedback; after these messages are classified, feedback is given, in the same order the messages are classified, using the same

standard interface as for the filtering task. The only apparent difference to the filter is that each classification request is not immediately followed by training for the classified message.

The exact sequence of classification requests and feedback is determined by a special index file supplied with the corpus. The delay intervals were selected at random using an exponential distribution with a mean corresponding to several day's delay – from 500 to 1000 messages, depending on the corpus. While the intervals were randomly generated, there was no variation in the presentation of feedback to the various filters; each filter saw exactly the same sequence.

Tools for training and development were supplied to participants in advance; index files specifying feedback delay were supplied for the the training corpora. Filters were evaluated on the same four corpora used for immediate feedback, augmented by index files with randomly generated delay.

It was anticipated that the delayed feedback task would be more difficult for the filters, and that filters might be able to harness information from unlabeled messages (one for which feedback had not yet occurred) to improve performance.

The track coordinator considered, in addition, the use of incomplete feedback in which the true classification for some messages was never communicated to the filter. While this scenario more closely models that of real filter deployment, we argue that this situation is aptly modeled by the task as deployed. Since the filter is always trained on past data and asked to classify future data, using incomplete judgements would simply be equivalent to using a smaller corpus of training data. Resource constraints limited the number of corpora we were able to use, and it was decided that the largest possible corpora would yield the highest statistical power.

2.3 The Active Learning Task

The active learning task models the situation in which a spam filter is first deployed. We assume that many historical email messages are available, but that these messages have not been adjudicated as ham or spam. The filter examines these messages as a batch (although it knows their chronology) and asks the user (or administrator) to adjudicate several before being deployed to filter new messages.

For the active learning task each corpus was divided chronologically in the ratio $9:1$. The (chronologically) first 90% of the messages were given to the filter without classification, while the last 90% were held back for testing. For $n = 100, 200, 400, 800, ...$ filters were allowed to query the true classification of n messages selected by the filter. Based on the results of these queries, the filters were then required to classify the test messages in sequence.

The learning task was effected by a *shell* program, written in C++, which read the entire corpus index (including gold standard judgements) and simulated n queries by examining the index. Filter classification and training were effected using the same interface as for the filtering tasks, but this interface was between the shell program and the participant filter, both of which were under control of the participant. A single run of the shell program was used to simulate, in succession, all values of $n \leq m$ where m is the number of messages in the corpus.

A standard shell which selected n messages at random was supplied as a baseline; participants were invited to modify the shell to use a better strategy, while adhering to the constraint that the labels for at most n messages should be used in classification. Training and development was effected on the same training corpora as for the filtering tasks; the same evaluation corpora were used as well.

3 Evaluation Measures

We used the same evaluation measures developed for TREC 2005. The tables and figures in this overview report Receiver Operating Characteristic (ROC) Curves, as well as $1 - ROCA(\%)$ – the area above the ROC curve, indicating the probability that a random spam message will receive a lower spamminess score than a random ham message.

The appendix contains detailed summary reports for each participant run, including ROC curves, 1-ROCA%, and the following statistics. The *ham misclassification percentage* ($hm\%$) is the fraction of all ham classified as spam; the *spam misclassification percentage* ($sm\%$) is the fraction of all spam classified as ham.

There is a natural tension between ham and spam misclassification percentages. A filter may improve one at the expense of the other. Most filters, either internally or externally, compute a spamminess score that

reflects the filter's estimate of the likelihood that a message is spam. This score is compared against some fixed threshold t to determine the ham/spam classification. Increasing t reduces $hm\%$ while increasing $sm\%$ and vice versa. Given the score for each message, it is possible to compute $sm\%$ as a function of $hm\%$ (that is, $sm\%$ when t is adjusted to as to achieve a specific $hm\%$) or vice versa. The graphical representation of this function is a Receiver Operating Characteristic (ROC) curve; alternatively a recall-fallout curve. The area under the ROC curve is a cumulative measure of the effectiveness of the filter over all possible values. ROC area also has a probabilistic interpretation: the probability that a random ham will receive a lower score than a random spam. For consistency with $hm\%$ and $sm\%$, which measure failure rather than effectiveness, spam track reports the area *above* the ROC curve, as a percentage ($(1 - ROCA)\%$). The appendix further reports $sm\%$ when the threshold is adjusted to achieve several specific levels of $hm\%$, and vice versa.

A single quality measure, based only on the filter's binary ham/spam classifications, is nonetheless desirable. To this end, the appendix reports *logistic average misclassification percentage* ($lam\%$) defined as $lam\% = logit^{-1}(\frac{logit(hm\%)+logit(sm\%)}{2})$ where $logit(x) = log(\frac{x}{100\%-x})$. That is, $lam\%$ is the geometric mean of the *odds* of ham and spam misclassification, converted back to a proportion[1]. This measure imposes no a priori relative importance on ham or spam misclassification, and rewards equally a fixed-factor improvement in the odds of either.

For each measure and each corpus, the appendix reports 95% confidence limits computed using a bootstrap method [2] under the assumption that the test corpus was randomly selected from some source population with the same characteristics.

4 Spam Filter Evaluation Tool Kit

All filter evaluations were performed using the *TREC Spam Filter Evaluation Toolkit*, developed for this purpose. The toolkit is free software and is readily portable.

Participants were required to provide filter implementations for Linux or Windows implementing five command-line operations mandated by the toolkit:

- **initialize** – creates any files or servers necessary for the operation of the filter
- **classify** *message* – returns ham/spam classification and spamminess score for *message*
- **train ham** *message* – informs filter of correct (ham) classification for previously classified *message*
- **train spam** *message* – informs filter of correct (spam) classification for previously classified *message*
- **finalize** – removes any files or servers created by the filter.

Track guidelines prohibited filters from using network resources, and constrained temporary disk storage (1 GB), RAM (1 GB), and run-time (2 sec/message, amortized). These limits were enforced incrementally, so that individual long-running filters were granted more than 2 seconds provided the overall average time was less than 2 second per query plus one minute to facilitate start-up.

The toolkit takes as input a test corpus consisting of a set of email messages, one per file, and an index file indicating the chronological sequence and gold-standard judgements for the messages. It calls on the filter to classify each message in turn, records the result, and at some time later (perhaps immediately) communicates the gold standard judgement to the filter.

Participants were supplied as well, with an active learning shell, *active.cpp*, which they modified to effect the active learning task.

The recorded results are post-processed by an evaluation component supplied with the toolkit. This component computes statistics, confidence intervals, and graphs summarizing the filter's performance.

5 Test Corpora

For TREC 2006, we used two public corpora, one English and one Chinese, as well as two private corpora derived from the same users' email as the TREC 2005 private corpora.

[1] For small values, odds and proportion are essentially equal. Therefore $lam\%$ shares much with the geometric mean average precision used in the robust track.

5.1 Public English Corpus – trec06p

Public Corpora				Private Corpora			
	Ham	Spam	Total		Ham	Spam	Total
trec06p	12910	24912	37822	MrX2	9039	40135	49174
trec06c	21766	42854	64620	SB2	9274	2695	11969
Total	34677	67766	102442	Total	18313	42830	61143

Table 1: Corpus Statistics

14000 ham messages were crawled from the web. Only messages with complete "Received from" headers were selected; the messages were ordered by the time and date on the first such header. These messages were adjudicated by human judges assisted by several spam filters – DMC [3] , Bogofilter and SpamProbe – using the methodology developed for TREC 21005. About 1000 spam messages were discovered in this set; 12910 were ham.

The 14000 crawled messages were augmented by approximately 38000 spam messages collected in May 2006. Each spam message was paired with a ham message. The header of the spam message was altered to make it appear to have been addressed to the same recipient and delivered to the same mail server during the same time frame as the ham message. "To:" and "From:" headers, as well as the message bodies, were altered to substitute names and email addresses consistent with the addressee. SpamProbe and Bogofilter were run on the corpora, and their dictionaries examined, to identify artifacts that might identify these messages. A handful were detected and removed; for example, incorrect uses of daylight saving time, and incorrect versions of server software in header information. The DMC spam filter was run on the corpus several times and disagreements between the filter and the recorded judgement were adjudicated.

5.2 Public Chinese Corpus - trec06c

The Public Chinese corpus used data provided by the CERNET Computer Emergency Response Team (CCERT) at Tsinghuaua University, Beijing. The ham messages consisted of those send by to a mailing list; the spam messages were those sent to a spam trap in the same internet domain. Headers and bodies of spam messages were modified to make them appear to have been delivered to the same servers as the ham messages, in the same time interval. Both the ham and spam messages were modified to as to remove structural elements not in common with those of the other class, such as embedded signature files, certain kinds of HTML markup, and the like.

Pilot filtering using DMC revealed that the Chinese corpus was quite easy to classify; it was felt nevertheless that the corpus would reveal any western bias in filtering strategies.

5.3 Private Corpus – MrX2

The MrX2 corpus was derived from the same source as the MrX corpus used for TREC 2005. For comparability with MrX, a random subset of X's email from October 2005 through April 2006 was selected so as to yield the same corpus size and ham/spam ratio as for MrX. This selection involved primarily the elimination of spam messages, whose volume had increased about 50% since the 2003-2004 interval in which the original MrX corpus was collection. Ham volume was insubstantially different.

5.4 Private Corpus – SB2

The SB2 corpus was collected from the same source as last year's SB corpus. Spam volume tripled since last year; all delivered messages were used in the corpus.

5.5 Aggregate Pseudo-Corpus

The subject filters were run separately on the various corpora. That is, each filter was subject to eight test runs – four with immediate feedback and four with delayed feedback. For each filter and each type of feedback, an *aggregate run* was created combining its results on the four corpora as if they were one. The evaluation component of the toolkit was run on the aggregate results, consisting of 163,641 messages

for each type of feedback – 52,989 spam and 110,652 ham. The summary results on the aggregate runs provide a composite view of the performance on all corpora, but are not the results of running the filter on an aggregate corpus; hence we dub the aggregate a pseudo-corpus.

6 Spam Track Participation

Group	Filter Prefix
Beijing University of Posts and Telecommunications	bpt
Harbin Institute of Technology	hit
Humboldt University Berlin & Strato AG	hub
Tufts University	tuf
Dalhousie University	dal
Jozef Stefan Institute	ijs
Tony Meyer	tam
Mitsubishi Electric Research Labs (CRM114)	CRM
Fidelis Assis	ofl

Table 2: Participant filters

Corpus / Task	Filter Suffix
trec06p / immediate feedback	pei
trec06p / delayed feedback	ped
trec06c / immediate feedback	pci
trec06c / delayed feedback	pcd
MrX2 / immediate feedback	x2
MrX2 /delayed feedback	x2d
SB2 / immediate feedback	b2
SB2 / delayed feedback	b2d

Table 3: Run-id suffixes

Nine groups participated in the TREC 2006 filtering tasks; five of them also participated in the active learning task. For each task, each participant submitted up to four filter implementations for evaluation on the private corpora; in addition, each participant ran the same filters on the public corpora, which were made available follwoing filter submission. All test runs are labelled with an identifier whose prefix indicates the group and filter priority and whose suffix indicates the corpus to which the filter is applied. Table 2 shows the identifier prefix for each submitted filter. All test runs have a suffix indicating the corpus and task, detailed in figure 3 .

7 Results

Figures 2 through 10 show the results for the filtering runs – immediate and delayed feedback – on the four corpora and on the aggregate pseudo-corpus. The left panel of each figure shows the ROC curve, while the right panel shows the learning curve: cumulative 1-ROCA% as a function of the number of messages processed. Only the best run for each participant is shown in the figures; table 13 shows 1-ROCA% for all filter runs on all corpora. Full details for all runs are given in the notebook appendix.

Figures 11 through 14 show the performance of the active learning filters as a function of n – the number of training messages. Only the best run from each participant is shown. Full details are given in the notebook appendix.

8 Conclusions

Although the Chinese corpus was much easier than the others, and SB2 was harder, results are generally consistent. With a few exceptions, performance on the delayed feedback task was inferior to that of the

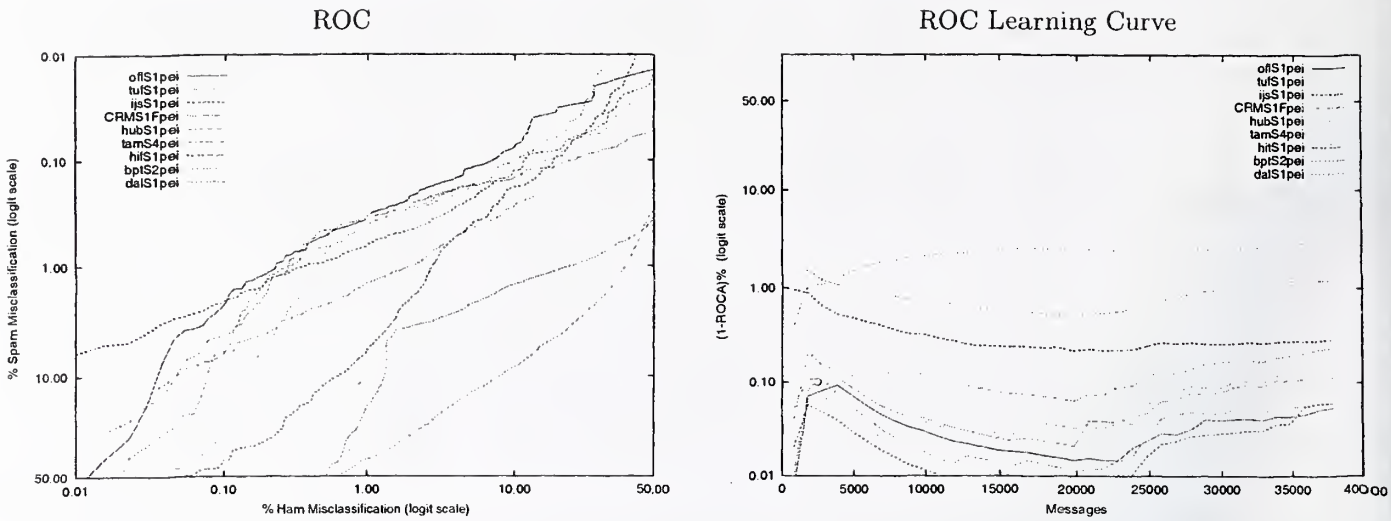


Figure 1: trec06p Public Corpus – Immediate Feedback

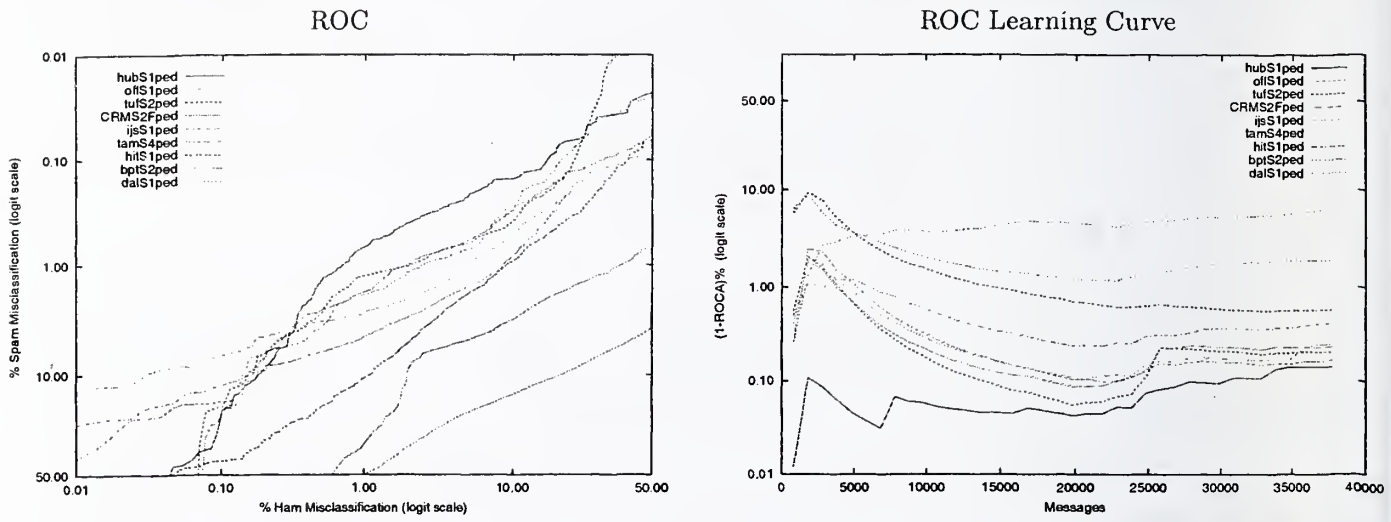


Figure 2: trec06p Public Corpus – Delayed Feedback

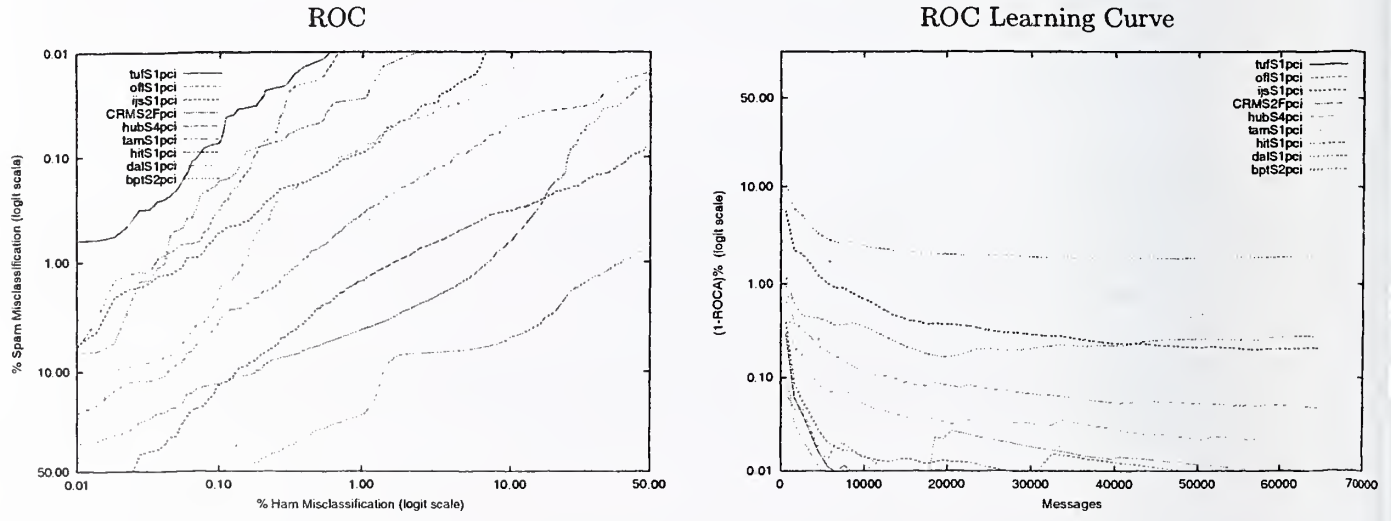


Figure 3: trec06c Chinese Corpus – Immediate Feedback

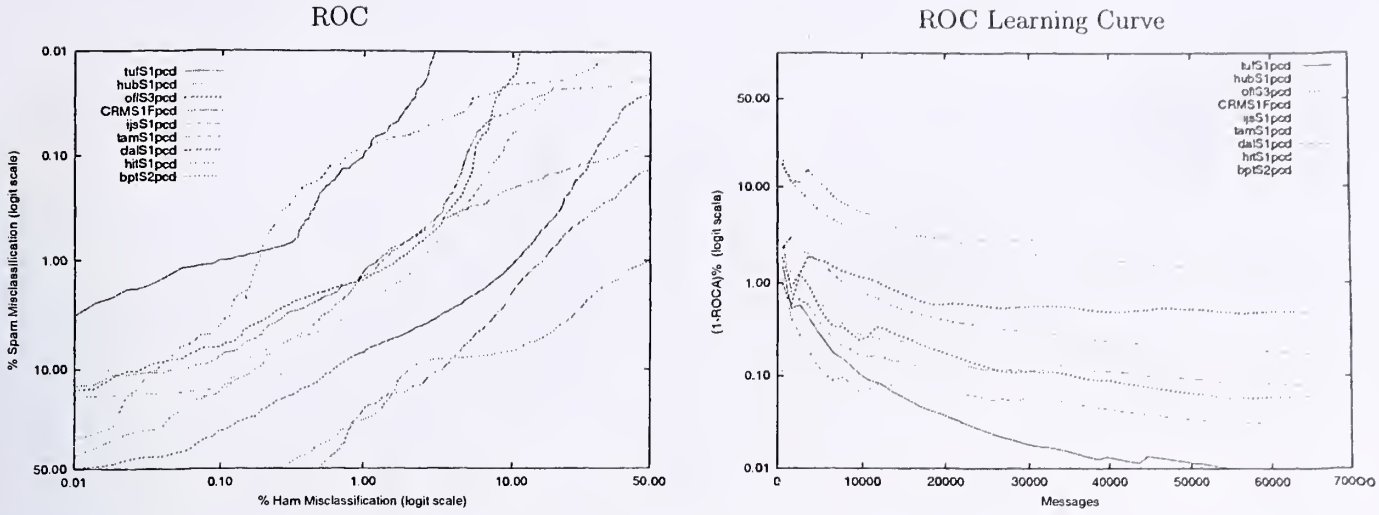


Figure 4: trec06c Chinese Corpus – Delayed Feedback

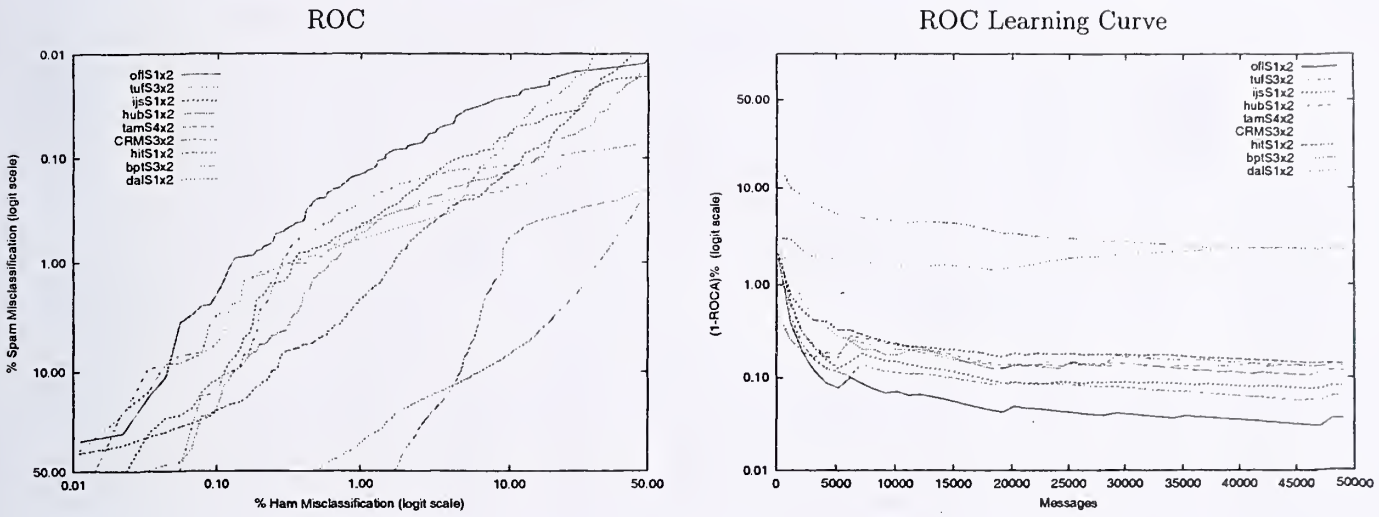


Figure 5: MrX2 Corpus – Immediate Feedback

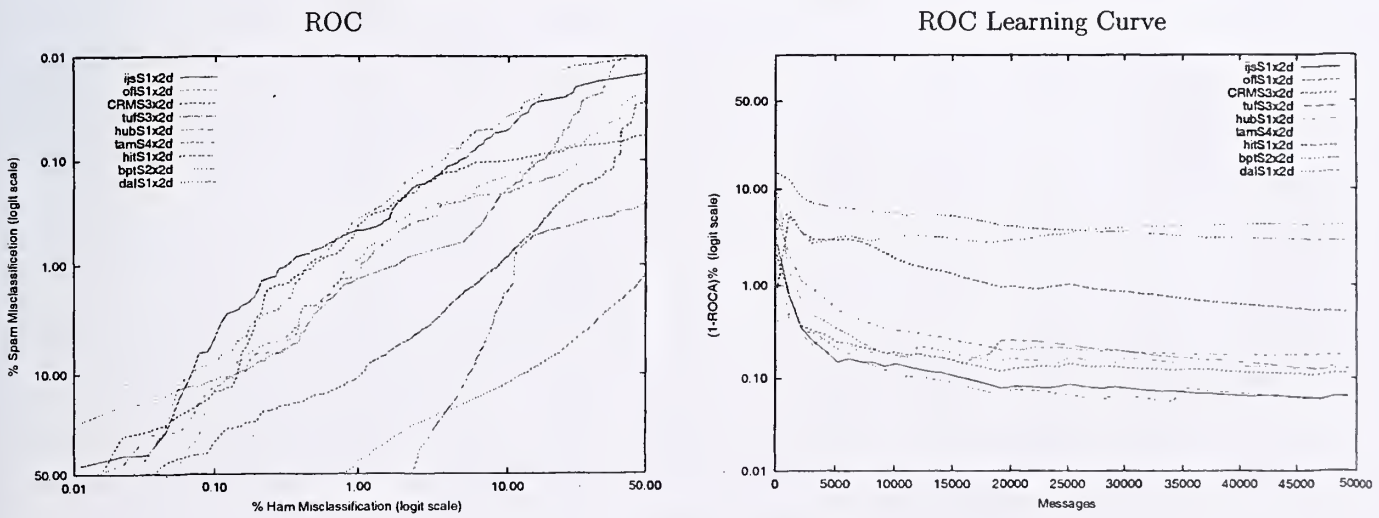


Figure 6: MrX2 Corpus – Delayed Feedback

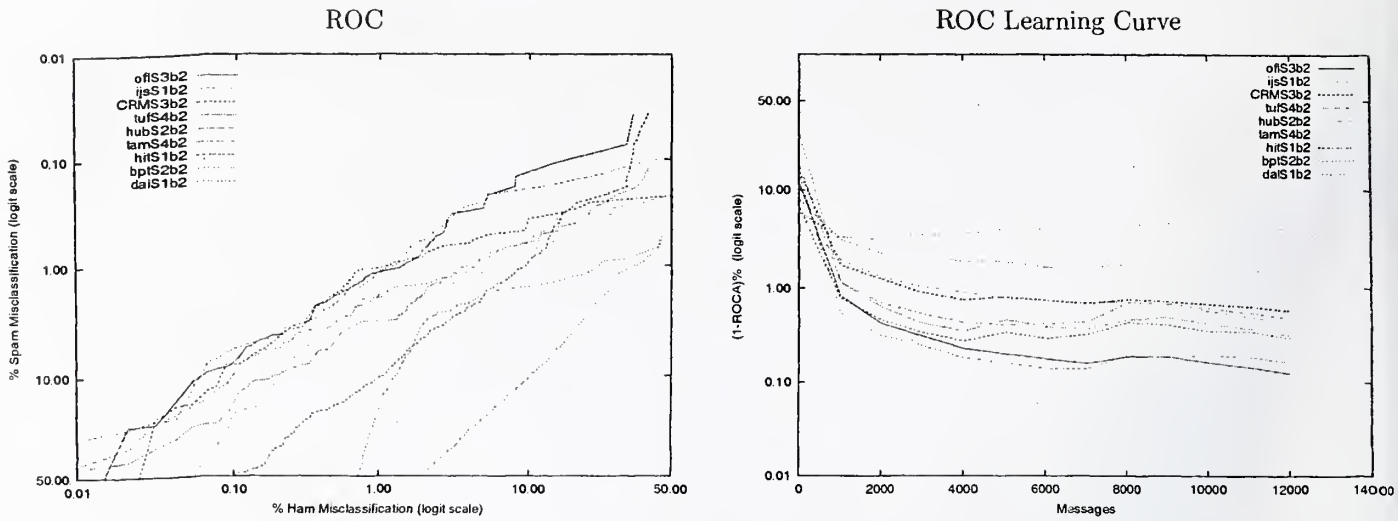


Figure 7: SB2 Corpus – Immediate Feedback

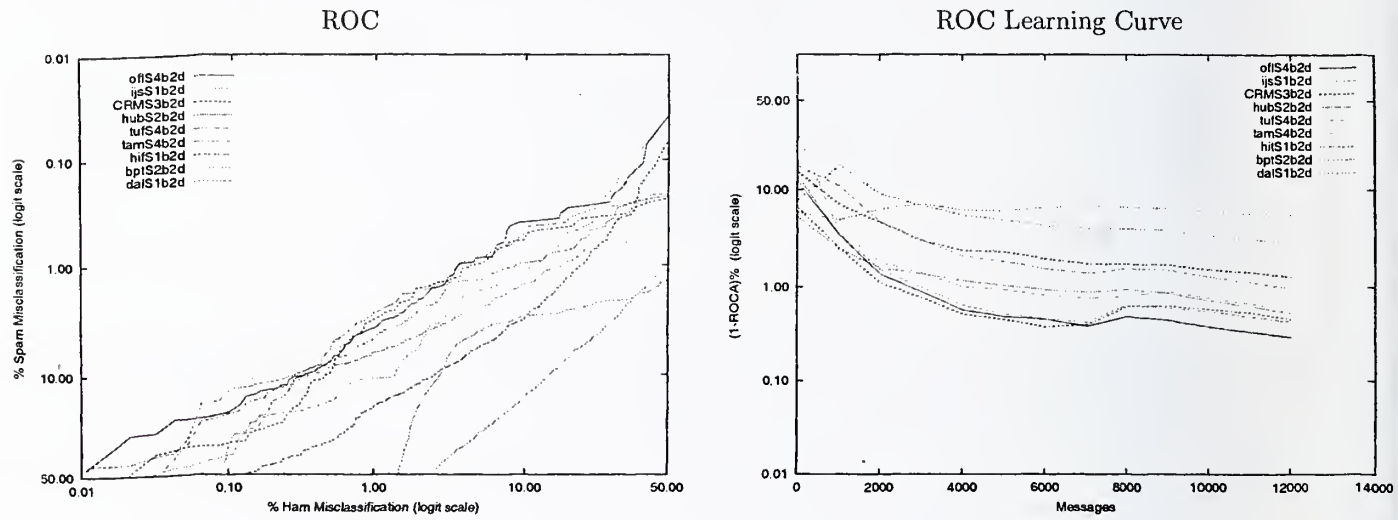


Figure 8: SB2 Corpus – Delayed Feedback

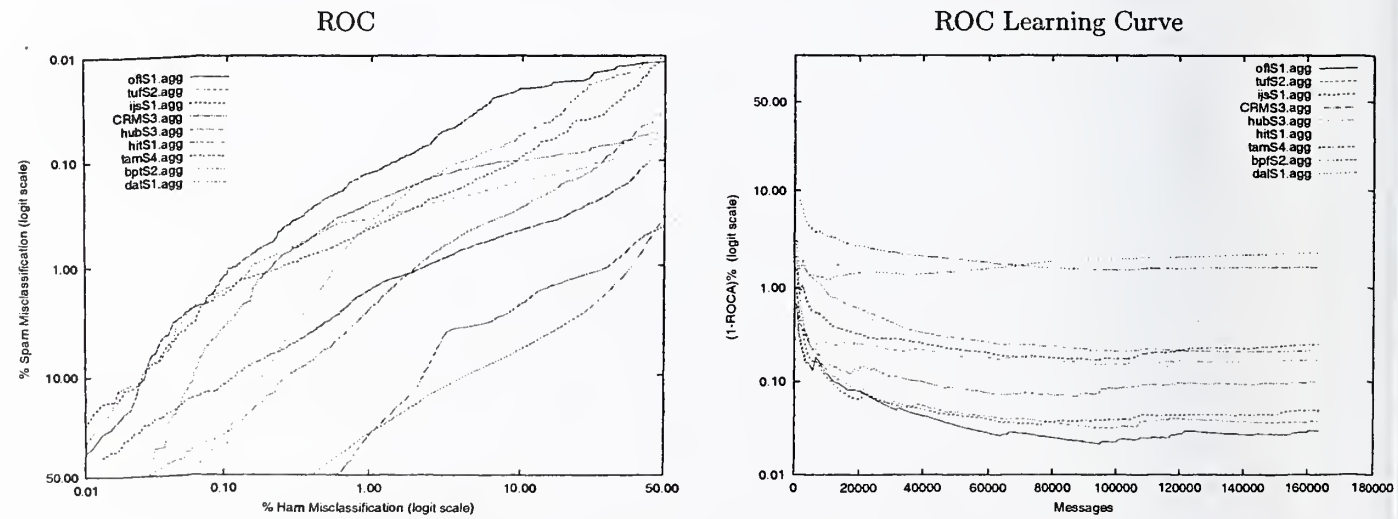
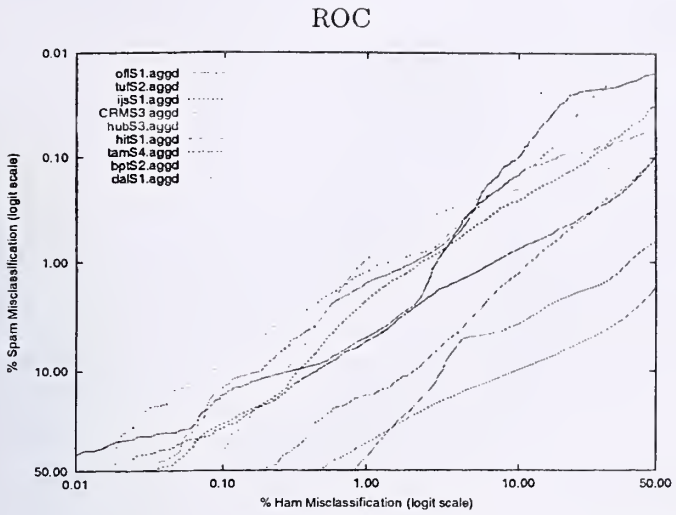


Figure 9: Aggregate Pseudo-Corpus – Immediate Feedback



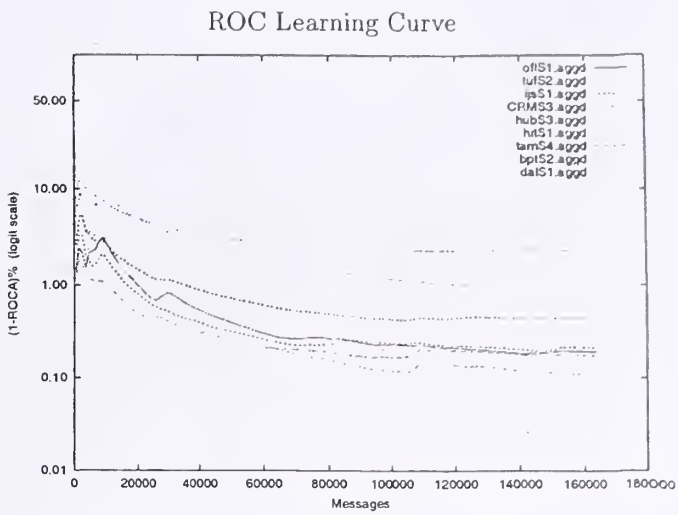


Figure 10: Aggregate Pseudo-corpus – Delayed Feedback

	Aggregate		trec06p		trec06c		MrX2		SB2	
Filter\Feedback	immediate	delay	immediate	delay	immediate	delay	immediate	delay	immediate	delay
oflS1	0.0295	0.1914	0.0540	0.1668	0.0035	0.0666	0.0363	0.0651	0.1300	0.3692
oflS3	0.0327	0.1908	0.0562	0.1702	0.0035	0.0601	0.0523	0.0824	0.1249	0.3174
oflS2	0.0365	0.2018	0.0597	0.2045	0.0104	0.1297	0.0525	0.0931	0.1479	0.3659
tufS2	0.0370	0.1079	0.0602	0.2038	0.0031	0.0104	0.0691	0.1449	0.3379	0.6923
oflS4	0.0381	0.1828	0.0583	0.1965	0.0077	0.0855	0.0718	0.1155	0.1407	0.2941
tufS1	0.0445	0.1262	0.0602	0.2110	0.0023	0.0081	0.0953	0.1991	0.3899	0.8361
ijsS1	0.0488	0.2119	0.0605	0.2457	0.0083	0.1117	0.0809	0.0633	0.1633	0.4276
tufS3	0.0705	0.1497	-	-	-	-	0.0633	0.1263	0.3350	0.6137
tufS4	0.0749	0.1452	-	-	-	-	0.0750	0.1314	0.3199	0.5696
CRMS3	0.0978	0.1743	0.1136	0.2762	0.0105	0.0888	0.1393	0.1129	0.2983	0.4584
CRMS2	0.1011	0.1667	0.1153	0.2325	0.0094	0.0975	0.1592	0.1143	0.4196	0.6006
CRMS1	0.1081	0.2165	0.1135	0.2447	0.0218	0.0784	0.1498	0.1341	0.3852	0.6346
hubS3	0.1674	0.2170	0.1564	0.1958	0.0353	0.0495	0.2102	0.2294	0.6225	0.8104
hubS4	0.1717	0.2400	0.1329	0.2006	0.0233	0.0330	0.1385	0.1763	0.5777	0.6784
hubS1	0.1731	0.2013	0.1310	0.1418	0.0238	0.0319	0.1180	0.1359	0.5295	0.5779
hubS2	0.1945	0.2716	0.1694	0.2952	0.0273	0.0369	0.1450	0.1827	0.4276	0.5306
hitS1	0.2112	0.8846	0.2884	0.5783	0.2054	1.3803	0.1412	0.5184	0.5806	1.2829
CRMS4	0.2375	1.5324	0.4675	2.1950	0.0579	1.7675	0.3056	0.4898	0.9653	2.0009
tamS4	0.2493	0.4480	0.2326	0.4129	0.1173	0.2705	0.1328	0.1755	0.4813	0.9653
tamS1	0.3008	1.0910	0.4103	0.8367	0.0473	0.1726	0.4011	0.6714	0.5912	4.5170
tamS2	0.9374	3.2366	1.2414	3.9352	0.4464	1.5370	-	-	6.5258	23.8125
tamS3	1.5309	2.2236	1.0602	1.8279	0.2899	1.0860	0.9514	1.5965	1.8462	6.0056
bptS2	1.6313	2.2999	1.2109	1.9264	1.8912	2.5444	2.5486	2.9571	1.4311	2.9050
bptS1	1.7867	2.6169	1.3690	2.0924	2.2829	3.0341	2.5926	3.6977	1.5545	2.9271
bptS3	1.9401	2.5669	1.3813	1.9520	2.9886	3.5715	2.3501	3.0866	1.6350	3.0487
bptS4	1.9818	2.6557	1.3215	1.9539	2.8267	3.3317	2.5100	3.4217	1.4970	3.0337
hitS2	2.1643	6.6776	0.8807	2.1074	3.2501	10.4413	1.2270	5.7253	1.9922	5.5975
dalS1	2.3278	4.0038	3.1383	6.3238	0.2739	0.4817	2.5035	4.3461	4.1620	5.6777
dalS2	3.2034	5.2315	4.7879	7.8412	0.4715	0.7934	5.8405	9.7809	6.9847	9.6615
hitS3	3.8063	7.8970	3.4365	7.9408	4.9442	9.6859	1.7927	5.5140	5.0801	7.7840
dalS3	6.2410	8.9847	4.0860	7.2674	0.7827	1.3635	22.9897	34.5147	4.6356	7.8237
dalS4	11.9983	19.3618	15.2705	28.3349	5.2031	9.4180	18.2197	25.6102	17.4061	24.1379

Table 4: Summary 1-ROCA (%)

baseline, as expected. It is not apparent that filters made much use of the unlabeled data in the delayed

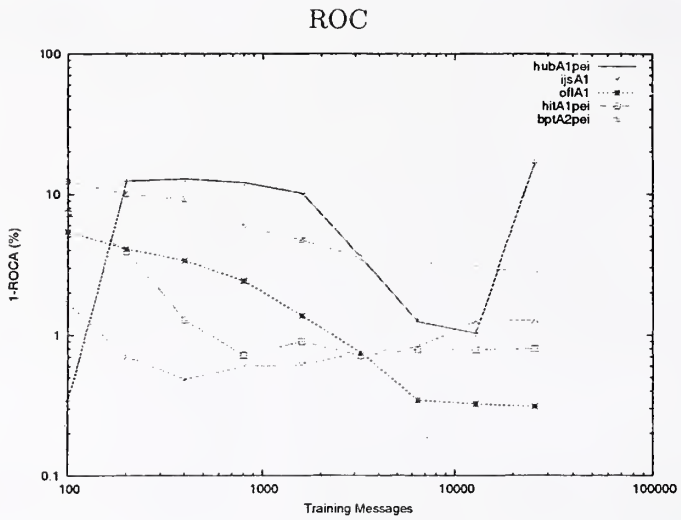


Figure 11: Active Learning – trec06p Public Corpus

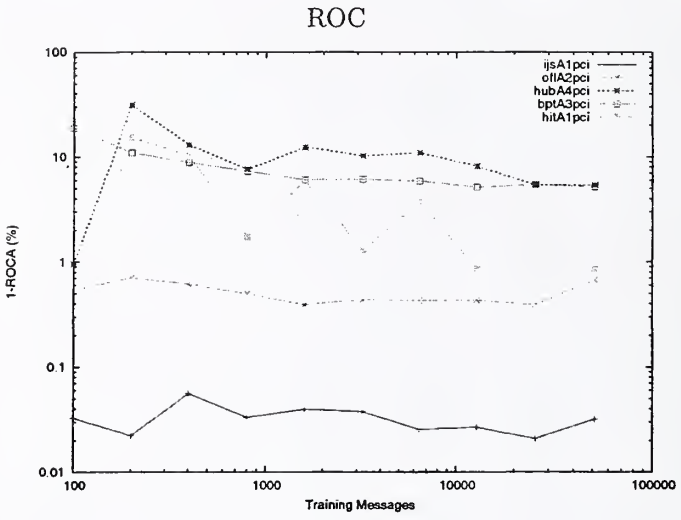


Figure 12: Active Learning – trec06c Chinese Corpus

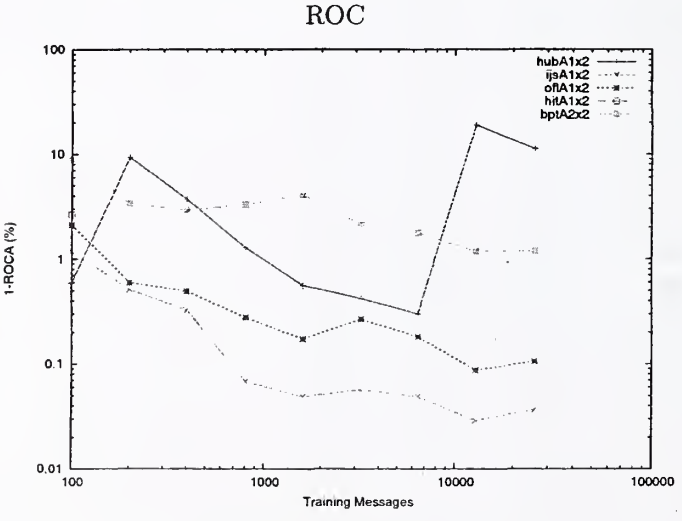


Figure 13: Active Learning – MrX2 Corpus

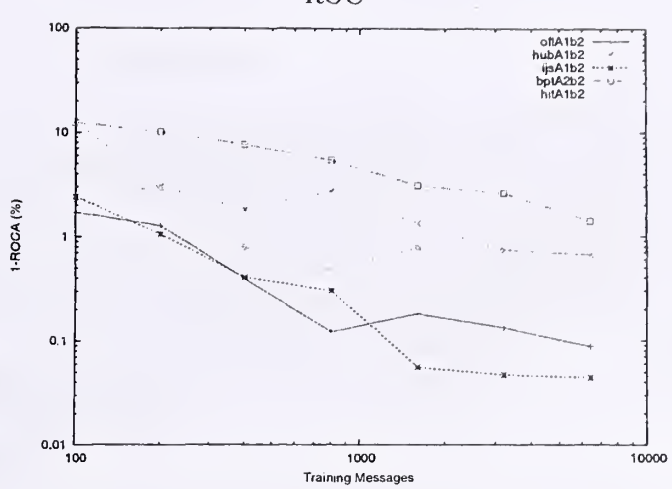


Figure 14: Active Learning – SB2 Corpus

feedback task; individual participant reports will reveal this. The active learning task presents a significant challenge.

Detailed comparison between TREC 2005 and TREC 2006 results have yet to be made, but it appears that

1. The best (and median) filter performance has improved over last year
2. The new corpora are no "harder" than the old ones: spammers have not defeated filters
3. Challenges remain in exploiting unlabeled data for spam classification, within the framework of the delayed filtering and active learning tasks.

9 Acknowledgements

The authors thank Stefan Buettcher and Quang-Anh Tran for their invaluable contributions to this effort.

References

[1] Cormack, G. Trec 2005 spam track overview. In *Proceedings of TREC 2005* (Gaithersburg, MD, 2005).

[2] Cormack, G. Statistical precision of information retrieval evaluation. In *Proceedings of SIGIR 2006* (Seattle, WA, 2006).

[3] Cormack, G., and Bratko, A. Batch and on-line spam filter evaluation. In *Proceedings of CEAS 2006* (Mountain View, CA, 2006).

The TREC 2006 Terabyte Track

Stefan Büttcher
University of Waterloo
stefan@buettcher.org

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

Ian Soboroff
NIST
ian.soboroff@nist.gov

1 Introduction

The primary goal of the Terabyte Track is to develop an evaluation methodology for terabyte-scale document collections. In addition, we are interested in efficiency and scalability issues, which can be studied more easily in the context of a larger collection.

TREC 2006 is the third year for the track. The track was introduced as part of TREC 2004, with a single adhoc retrieval task. For TREC 2005, the track was expanded with two optional tasks: a named page finding task and an efficiency task. These three tasks were continued in 2006, with 20 groups submitting runs to the adhoc retrieval task, 11 groups submitting runs to the named page finding task, and 8 groups submitting runs to the efficiency task. This report provides an overview of each task, summarizes the results, and outlines directions for the future. Further background information on the development of the track can be found in the 2004 and 2005 track reports [4, 5].

For TREC 2006, we made the following major changes to the tasks:

1. We strongly encouraged the submission of adhoc manual runs, as well as runs using pseudo-relevance feedback and other query expansion techniques. Our goal was to increase the diversity of the judging pools in order to a create a more re-usable test collection. Special recognition (and a prize) was offered to the group submitting the run contributing the most unique relevant documents to the judging pool.

2. The named page finding topics were created by task participants, with each group asked to create at least 12 topics.

3. The experimental procedure for the efficiency track was re-defined to permit more realistic intra- and inter-system comparisons, and to generate separate measurements of latency and throughput. In order to compare systems across various hardware configurations, comparative runs using publicly available search engines were encouraged.

2 The Document Collection

All tasks in the track use a collection of Web data crawled from Web sites in the gov domain during early 2004. We believe this collection ("GOV2") contains a large proportion of the crawlable pages present in gov at that time, including HTML and text, along with the extracted contents of PDF, Word and postscript files. The collection is 426GB in size and contains 25 million documents. For TREC 2004, the collection was distributed by CSIRO, Australia, who assisted in its creation. For 2005 and 2006, the collection was distributed by the University of Glasgow.

3 Adhoc Retrieval Task

3.1 Task Description

An adhoc retrieval task investigates the performance of systems that search a static set of documents using previously-unseen topics. For each topic, participants create a query and generate a ranked list of documents. For the 2006 task, NIST created and assessed 50 new topics. An example is provided in Figure 1. In addition, the 99 topics created in 2004 and 2005 (topics 701-800) were re-used for automatic runs.

As is the case for most TREC adhoc tasks, a topic describes the underlying information need in several forms. The title field essentially contains a keyword query, similar to a query that might be entered into a Web search engine. The description field provides a longer statement of the topic requirements, in the form of a complete sentence or question. The narrative, which may be a full paragraph in length, supplements the other two fields and provides additional information required to specify the nature of a relevant document.

For the adhoc task, an experimental run consisted of the top 10,000 documents for each topic. To generate a run, participants could create queries automatically or manually from the topics. For a run to be considered *automatic* it must be created from the topics without any human intervention. All other runs are manual. Manual runs used only the 50 new topics; automatic runs used all 149 topics from 2004–2006.

For most experimental runs, participants could use any or all of the topic fields when creating queries from the topic statements. However, a group submitting any automatic run was required to submit at least one automatic run that used only the title field of the topic statement. Manual runs were encouraged, since these runs often add relevant documents to the evaluation pool that are not found by automatic systems using current technology. We offered a prize to the group with the run that returned the most unique relevant documents. Groups could submit up to five runs.

Runs were pooled by NIST for judging. The details of the pooling process differ substantially from previous years, and from previous TREC adhoc tasks, and are detailed in a separate section.

Assessors used a three-way scale of "not relevant", "relevant", and "highly relevant". A document is considered relevant if any part of the document contains information which the assessor would include in a report on the topic. It is not sufficient for a document to contain a link that appears to point to a relevant Web page, the document itself must contain the relevant information. It was left to the individual assessors to determine their own criteria for distinguishing between relevant and highly relevant documents. For the purpose of computing effectiveness measures that require binary relevance judgments, the relevant and highly relevant documents are combined into a single "relevant" set.

```
<top>
<num> Number: 835

<title> Big Dig pork

<desc> Description:
Why is Boston's Central Artery project, also known as "The Big Dig",
characterized as "pork"?

<narr> Narrative:
Relevant documents discuss the Big Dig project, Boston's Central
Artery Highway project, as being a big rip-off to American taxpayers
or refer to the project as "pork". Not relevant are documents which
report fraudulent acts by individual contractors. Also not relevant
are reports of cost-overruns on their own.

</top>
```

Figure 1: *Adhoc Task Topic 835*

In addition to the top 10,000 documents for each run, we collected details about the hardware and software configuration used to generate them, including performance measurements such as total query processing time. For total query processing time, groups were asked to report the time required to return the top 20 documents, not the time to return the top 10,000. It was acceptable to execute a system twice for each run, once to generate the top 10,000 documents and once to measure the execution time for the top 20 documents, provided that the top 20 documents were the same in both cases.

3.2 Adhoc Pooling

Last year, we reported that the pools were increasingly dominated by documents containing the title words of topics, to the possible exclusion of other relevant documents [5, 6, 2]. This could lead to a test collection that is biased towards simple title-query-based retrieval approaches and that may not measure other systems fairly. We were able to observe this phenomenon directly in the AQUAINT collection due to the presence of a feedback run that retrieved many unique relevant documents. It seemed clear that it must also occur in GOV2, although we had no direct evidence in the form of missed relevant documents. For 2006, we took steps to gather data to answer the bias question for GOV2, including the active solicitation of manual runs in an effort to diversify the pools.

In addition, we required the adhoc runs to include the older topics to determine if the newer runs retrieved unusual amounts of unjudged documents for these topics. If so, it would provide more evidence of bias in the collection as well as data for analyzing the impact of that bias. However, since this activity was limited to automatic runs, we did not expect to see these runs fall far from the original qrels.

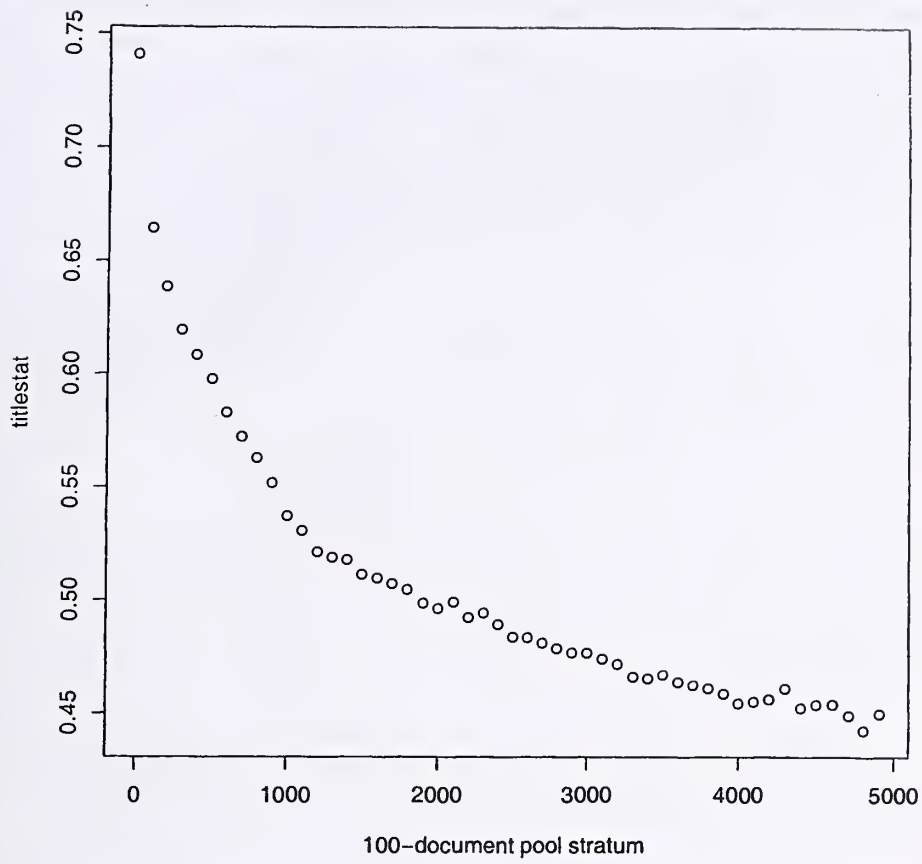


Figure 2: *Values for titlestat in 100-document strata of the pool.*

Lastly, we constructed three separate pools for the 2006 topics, two of which were used in the evaluation. The runs contributing to the pools are the same for all three: one automatic, one manual, and one efficiency task run per group. If a group only did manual or automatic runs, we took two of that type from that group.

The first pool is a traditional TREC pool to depth 50. This pool allows us to report traditional measures such as MAP and precision at fixed document cutoffs to some degree of accuracy, and thus provide some degree of continuity to track participants while experimental methods were tried.

The second pool is also a traditional TREC pool, but drawn starting at depth 400. This pool was motivated by the values of the *titlestat* measure, described by Buckley et al. [2]. The authors of that paper primarily computed the titlestat of judged relevant documents, but titlestat can be computed over any group of documents. For this application, we computed the titlestat of 100-document strata of the "complete" pool (that is, all documents pooled to depth 10,000). These titlestats are plotted to depth 5000 in figure 2. Whereas the pool from depths 1-100 has a titlestat of 0.74, at depths 400-500 the titlestat of the pool is just over 0.6. This pool starts at depth 400, but goes down to a different depth for each topic, such that the total size per topic for this plus the initial depth-50 pool is 1000 documents. While this pool was not used for the evaluation, the

relevance judgments allow us to see if relevant documents still occur frequently in lower-titlestat regions.

The third pool is a random sample, drawn in such a way as to attempt to gather more relevant documents deeply from topics where we expect them to occur. Using the relevance judgments from the depth-50 pool, we calculate the probability of relevance given the pool rank of a document. Using a simple linear fit based on experiments with last year's pools, we then estimate the depth at which we will find a given ratio of relevant documents to pool size. We then draw a random sample of about 200 documents up to that target depth. This third pool varies in maximum pool depth from 57 to 1252 depending on how many relevant documents were found in the depth-50 pool for each topic.

The qrels from the third pool are not usable with MAP and other traditional measures. Instead, they are intended to be used with a new measure called *inferred average precision*, or *infAP* [7].[1] infAP estimates average precision based on a known pool where only a sample of documents in the pool are judged. The expected precision at rank k is

$$\frac{1}{k} \cdot 1 + \frac{k-1}{k} \left(\frac{P}{k-1} \cdot \frac{R+\epsilon}{R+N+\epsilon} \right)$$

where P is the number of documents in the pool, R is the number of known relevant documents above rank k, and N is the number of known nonrelevant documents above rank k. infAP is the average of these precisions at each relevant document. infAP is similar to bpref in that it is intended for incomplete judgments but differs in that it is an direct estimate of average precision based on a sample.

The official evaluation results report MAP (and other standard trec_eval measures) on the depth-50 pool, infAP on the random-sample pool, and for automatic runs MAP on all 149 terabyte track topics (where the AP scores for the 2006 topics are from the depth-50 pool.)

3.3 Adhoc Results

Table 1 provides an summary of the results obtained on the title-only automatic runs sorted by bpref. Only the best run from each group is shown. Figure 2 provides the same information for the manual runs. The first two columns of each table identify the group and run. The next three columns provide the values of three standard effective measures for each run: bpref [3], precision at 20 documents (p@20), and mean average precision (MAP). The sixth colunm provides values for the new infAP measure decribed above. The last two columns list the number of CPUs used to generate the run and the total query processing time.

The top-scoring automatics runs were generated using various retrieval methods, including Okapi BM25 and language modeling approaches. Many of these runs took features such as phrases, term proximity and matches in the title field into account during ranking. Of particular note is the prevalence of pseudo-relevance feedback, which substantially improved performance for most groups. On the other hand, none of the top-eight runs used anchor text, and only one used link analysis techniques.

A prize was offered to the group submitting the run containing the most unique relevant documents, excluding other runs from the same group. The prize (a NIST clock) was awarded to Chris Buckley of Sabir Research for the run sabtb06man1, which contributed 222 unique documents.

[1] A technical note describing infAP can be found at `http://trec.nist.gov/act_part/tracks/terabyte/inferredAP.pdf`

Group	Run	bpref	p@20	MAP	infAP	CPUs	Time (sec)
uwaterloo-clarke	uwmtFadTPFB	0.4251	0.5570	0.3392	0.2999	1	964
umass.allan	indri06AlceB	0.4229	0.5410	0.3687	0.3157	1	38737
pekingu.yan	TWTB06AD01	0.4193	0.5150	0.3737	0.3224	4	56160
hummingbird.tomlinson	humT06xle	0.4172	0.5820	0.3452	0.2947	1	36000
ibm.carmel	JuruTWE	0.4002	0.5670	0.3506	0.2687	1	3375
uglasgow.ounis	uogTB06QET2	0.3995	0.5400	0.3456	0.2861	1	N/A
ecole-des-mines.beigbeder	AMRIMtp20006	0.3942	0.5170	0.3120	0.2994	2	68344
coveo.soucy	CoveoRun1	0.3886	0.5440	0.3296	0.2564	5	135
umilano.vigna	mg4jAutoV	0.3774	0.4510	0.2882	0.2765	4	3000
rmit.scholer	zetadir	0.3726	0.4800	0.3056	0.2599	2	466.5
umelbourne.ngoc-anh	MU06TBa2	0.3682	0.5130	0.3039	0.2549	1	25.25
uamsterdam.ilps	UAmsT06aTeLM	0.3528	0.4850	0.2958	0.2363	2	2394
dublincityu.gurrin	DCU05BASE	0.3509	0.5090	0.2695	0.2067	1	495
tsinghuau.zhang	THUADALL	0.3432	0.4600	0.2858	0.2444	4	560
lowlands-team.deVries	CWI06DISK1ah	0.3361	0.4780	0.2770	0.2299	1	60.3
polytechnicu.suel	p6tbadt	0.3073	0.3920	0.2274	0.1972	1	60
max-planck.theobald	mpiirtitle	0.2849	0.4270	0.1805	0.1678	2	38
northeasternu.aslam	hedge0	0.2568	0.3460	0.1771	0.1388	4	110000
sabir.buckley	sabtb06at1	0.2434	0.3250	0.1361	0.1045	1	77
ualaska.fairbanks.newby	arscDomAlog	0.1463	0.0550	0.0541	0.0675	108	120000

Table 1: *Adhoc Results. Best automatic title-only run from each group, according to bpref.*

4 Named Page Finding Task

4.1 Task Description

Users sometimes search for a page by name. In such cases, an effective search system will return that page at or near rank one. In many cases there is only one correct answer. In other cases, any document from a small set of "near duplicates" is correct. The objective of the task, therefore, is to find a particular page in the GOV2 collection, given a topic that describes it. For example, the query "`Apollo 11 Mission`" would be satisfied by NASA's history page on the first moon landing.

Named page topics were created by track participants through a purpose-built Web interface to the Wumpus search engine[2] and hosted at the University of Waterloo. Participants were asked to imagine they were using a search engine to locate an interesting page that they found once but couldn't quite remember where it was. Their goal was to identify interesting, "bookmark-worthy" pages, that they thought they might want to go back and find again. Once such a page was found, they were to give it a name such as they might assign to a bookmark. The name was to approximate what they might type into a Web search engine to locate that page again.

Participants could identify interesting pages in one of two ways. One was to request random pages from the search engine, and to keep looking at random pages until one struck their fancy.

[2] `http://www.wumpus-search.org/`

Group	Run	bpref	p@20	MAP	infAP	CPUs	Time (sec)
uwaterloo-clarke	uwmtFmanual	0.4785	0.7030	0.4246	0.3503	1	20000
sabir.buckley	sabtb06man1	0.4104	0.6070	0.2666	0.2161	1	21600
pekingu.yan	TWTB06AD02	0.4089	0.5070	0.3152	0.2749	4	9625
rmit.scholer	zetaman	0.3976	0.5290	0.2873	0.2369	2	307
umilano.vigna	mg4jAdhocBV	0.3944	0.4930	0.2822	0.2465	4	610
umelbourne.ngoc-anh	MU06TBa1	0.3900	0.5420	0.2927	0.2431	8	15.50
ecole-des-mines.beigbeder	AMRIMtpm5006	0.3793	0.4390	0.2705	0.2702	2	39032
ibm.carmel	JuruMan	0.3570	0.5190	0.2754	0.2410	1	60
northeasternu.aslam	hedge30	0.3180	0.5110	0.2561	0.1942	4	18000
max-planck.theobald	mpiirmanual	0.3041	0.4810	0.1981	0.1692	2	25
ualaska.fairbanks.newby	arscDomManL	0.1202	0.0400	0.0351	0.0511	108	150000

Table 2: *Adhoc Results (manual runs), sorted by bpref.*

Another was to search for subjects of interest to the participant, and to look through the search results until something worth keeping was found.

Participants were instructed to make each page's name specific to that page. To check this, they were requested to perform a search with the name as a query, and to check to see if other pages came up which could take the same name. The named page itself did not need to appear in these search results, although it was acceptable if it did. The purpose of this check search was to weed out similar (but not near-duplicate) pages that might need to be distinguished in order to obtain a good named page topic. Near-duplicates of the page, which differ only in formatting or by trivial content changes, were permitted.

When evaluating the submitted runs, we identified these near-duplicates using NIST's implementation of Bernstein and Zobel's DECO algorithm [1]. We ran DECO on the top 100 retrieved documents from all submitted named page runs, identified near-duplicates of the known targets, and manually checked those for relevance. Near-duplicates are treated as equivalent to the original page and are included in the qrels file.

4.2 Named Page Finding Results

Figure 3 summarizes the results of the named page finding task. The performance of the runs is evaluated using three metrics:

- **MRR**: The mean reciprocal rank of the first correct answer.
- **% Top 10**: The proportion of queries for which a correct answer was found in the first 10 search results.
- **% Not Found**: The proportion of queries for which no correct answer was found in the results list.

The figure lists the best run from the each group by MRR. In addition, the figure indicates the runs that exploit link analysis techniques (such as pagerank), anchor text, and document structure (such as giving greater weight to terms appearing in titles).

Group	Run	MRR	% Top 10	% Not Found	CPUs	Time (sec)	Links?	Anchors?	Structure?
umass.allan	indri06Nsdp	0.512	69.6	13.8	6	10860	Y	Y	Y
uglasgow.ounis	uogTB06MP	0.466	65.2	12.7	1		N	Y	Y
coveo.soucy	CoveoNPRun2	0.431	59.1	19.9	5	235	N	N	Y
tsinghuau.zhang	THUNPNOSTOP	0.430	64.1	16.0	2	3240	N	Y	N
hummingbird.tomlinson	humTN06dpl	0.408	56.9	13.3	1	4600	N	N	Y
umelbourne.ngoc-anh	MU06TBn5	0.397	62.4	13.8	1	50	N	Y	N
rmit.scholer	zetnpft	0.389	54.7	19.3	2	2001	N	N	Y
uwaterloo-clarke	uwmtFnpsRR1	0.386	54.7	18.8	1	1149	Y	Y	Y
uamsterdam.ilps	UAmsT06n3SUM	0.363	55.2	23.8	2	2545	N	Y	Y
cas-ict.wang	icttb0603	0.337	44.2	28.7	1	427	N	N	Y
pekingu.yan	TWTB06NP02	0.238	34.3	44.2	4	3240	N	N	Y

Table 3: *Named Page Finding Results. Best run from each group, according to MRR.*

5 Efficiency Task

5.1 Task Description

The efficiency task extends both the adhoc task and the named page finding task, providing a vehicle for discussing and comparing efficiency and scalability issues in IR systems by defining better methodology to determine query processing times.

Two weeks before the new topics for the adhoc task were made available, NIST released a set of 100,000 efficiency topics. These topics were extracted from the logs of a commercial Web search engine. Because an analysis of last year's 50,000 efficiency topics, which also had been extracted from a Web search engine log, had shown that the topics did not match GOV2 very well and consequently could be processed much faster than the adhoc topics, this year we made sure that each query in the efficiency topic set:

- had produced at least one clickthrough to .gov in the Web search engine, and
- matched at least 20 documents in GOV2 (Boolean OR).

After creating a set of representative topics in this way, the title fields of the adhoc topics (751–850) and the named page finding topics (NP601–NP872, NP901–NP1081) from this year's and last year's Terabyte track were seeded into the topic set, but were not distinguished in any way. Figure 3 provides some examples from the resulting topic set. Participating groups were required to process these topics automatically; manual runs were not permitted.

The efficiency topic set was distributed in 4 separate files, representing 4 independent query streams. Groups were required to process queries within the same stream sequentially and in the order in which they appear in the topic file. Processing of each query in a stream was to be

```
68964:easy to do science projects for 5th grade
68965:who to contact when civil rights are violated
68966:ergonomic courses illinois
68967:big dig pork
68968:signs of a partial seizure
68969:food at jfk
68970:natural gas power point
68971:va home care portland oregon
68972:lexapro package insert
```

Figure 3: *Efficiency Task Topics 68964 to 68972*

completed before processing of the next query was started. Queries from different query streams could be processed concurrently or interleaved in any arbitrary order. The existence of independent query streams allowed systems to take better advantage of parallelism and I/O scheduling.

Each participating group ran their system on the entire topic set (all four streams), reporting the top 20 documents for each topic, the average per-topic reponse time (referred to as query processing *latency*), and the total time between reading the first topic and writing the last result set (used to calculate query throughput). The total time was recorded without taking into account system startup times. By processing all queries strictly sequentially, latency was minimized. A group was able to choose, however, to process queries from different streams in parallel in order to make better use of parallelism and to increase their system's query throughput.

In general, document retrieval systems can employ two different kinds of parallelism: intra-query and inter-query. With intra-query parallelism, the same query is processed on multiple CPUs in parallel, for example by splitting the document collection into equal parts and distributing these parts among different machines. Intra-query parallelism improves both latency and throughput, although not necessarily in a linear fashion. Inter-query parallelism, on the other hand, refers to the situation in which multiple queries are being processed at the same time. It can be used to increase query throughput, usually in a linear or near-linear way, but does not improve query latency. Distributing the 100,000 efficiency topics in four independent streams was meant to explicitly encourage inter-query parallelism and to allow groups to study latency/throughput trade-offs.

One of the goals of the Terabyte track is to be able to compare different approaches to high-performance information retrieval and to evaluate them quantitatively. The validity of direct comparisons between groups, however, is limited by the range of hardware used, which varies from desktop PCs to supercomputers. Last year, we tried to overcome this issue by applying some informal normalizations, based on the number of CPUs and the total cost of a system. Those attempts were only partially successful. In order to obtain more reliable inter-group performance comparisons, participants this year were were encouraged to submit at least one run that was conducted in a single-CPU configuration, with all queries being processed sequentially. They were also encouraged to download special TREC versions of the three open-source information retrieval systems

- Indri (`http://www.lemurproject.org/indri/`),
- Wumpus (`http://www.wumpus-search.org/`), and
- Zettair (`http://www.seg.rmit.edu.au/zettair/`);

Run	Number of CPUs	System cost (USD)	Query streams	Latency			Throughput			Effectiveness	
				measured (ms/query)	CPU-normalized	cost-normalized	measured (queries/s)	CPU-normalized	cost-normalized	P@20 (801–850)	MRR (901–1081)
CWI06DIST8	16	6,400	4	13	211	85	185.5	11.6	29.0	0.4680	0.181
CWI06MEM4	4	10,000	4	80	322	805	48.7	12.2	4.9	0.4720	0.190
humTE06i3	1	5,000	1	1,680	1,680	8,400	0.6	0.6	0.1	0.3690	0.123
humTE06v2	1	5,000	1	4,630	4,630	23,150	0.2	0.2	0.0	0.4290	0.373
mpiiotopk2p	2	5,000	4	29	57	143	35.0	17.5	7.0	0.4330	0.280
mpiiotopkpar	2	5,000	4	74	148	369	13.6	6.8	2.7	0.4280	0.291
MU06TBy6	1	500	4	55	55	28	18.2	18.2	36.4	0.4890	0.271
MU06TBy2	1	500	1	229	229	114	4.4	4.4	8.8	0.5050	0.256
p6tbep8	1	1,400	1	109	109	153	9.1	9.1	6.5	0.3890	0.254
p6tbeb	1	1,400	1	167	167	234	6.0	6.0	4.3	0.4540	0.244
rmit06effic	2	4,000	1	2,202	4,404	8,808	0.5	0.2	0.1	0.4650	0.258
THUTeraEff02	4	2,000	1	534	2,136	1,068	1.9	0.5	0.9	0.1500	0.222
THUTeraEff01	4	2,000	1	808	3.232	1,616	1.2	0.3	0.6	0.3920	0.246
uwmtFdcp03	1	1,800	1	13	13	23	80.0	80.0	44.4	0.4110	0.164
uwmtFdcp12	1	1,800	1	32	32	58	31.4	31.4	17.4	0.4790	0.219

Table 4: *Efficiency Results. Best run (according to P@20) and fastest run (according to average query latency) from each participating group. Effectiveness is reported for the the adhoc (P@20) and the named page finding (MRR) topics from 2006, not for the 2005 topics also present in the efficiency topics.*

to compile them, build an index for GOV2, and run 10,000 queries (a subset of the 100,000 efficiency topics) through these systems. For each such comparative efficiency run, participants were required to report the total time to process all queries and the total CPU consumed, i.e., the time during which the CPU was busy processing queries and not waiting for the hard drive, for instance.

Our incentive was that this data could be used to find out whether an inter-group performance comparison, across different hardware configurations, is feasible at all and also how efficiency numbers need to be normalized in order to obtain a fair comparison of two systems running on different hardware.

We would like to point out that the versions of the three systems we used for comparative reasons were modified and re-packaged for this specific purpose, without special support from their developers. It is therefore unlikely that the efficiency numbers that were obtained reflect the true performance of the systems. They were used exclusively to determine the performance of the hardware they were run on.

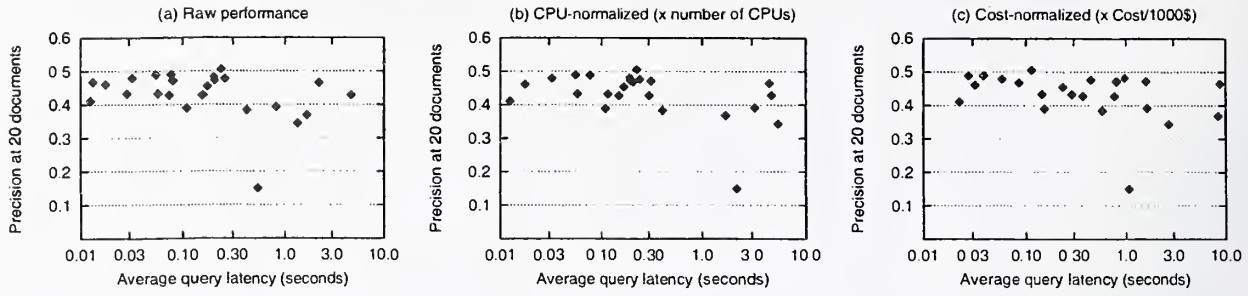


Figure 4: *Efficiency (average latency for the 100,000 efficiency topics) and effectiveness (P@20 for adhoc topics 801–850) of all 25 runs submitted for the efficiency task.*

5.2 Efficiency Results and Discussion

The efficiency results for the 100,000 efficiency topics are shown in Table 4. For each group, the best run (according to mean precision at 20 documents, for adhoc topics 801–850) and the fastest run (according to average query latency) are summarized. Both performance and precision vary greatly among the systems. The two fastest runs exhibit an average latency of 13 ms per query, while the slowest run consumed almost 5 seconds. P@20, on the other hand, varies quita substantially as well, between 0.150 and 0.505.

Although groups were allowed to have their system process queries from the four separate query streams in parallel, most groups chose not to do so. Only 5 out of the 25 efficiency runs processed queries in parallel. For some of these runs, however, the gains achieved from processing queries in parallel are tremendous, leading to a query throughput of up to 185 queries per second in the case of CWI06DIST8.

The efficiency measures reported in Table 4 include throughput and latency, both in their raw form (as measured by the respective group) as well as normalized, applying the same ad-hoc performance normalizations (CPU-normalized and cost-normalized) as last year:

- the CPU-normalized query latency is the real query latency, multiplied by the total number of CPUs in the system;
- the cost-normalized query latency compares each run with a run conducted on a hypothetical computer system costing 1,000 USD; thus, the latency of a run conducted on a $5,000 computer was multiplied by 5.

Efficiency vs. effectiveness trade-off plots, both normalized and non-normalized, for all 25 efficiency runs are shown in Figure 4.

The validity of calculating CPU-normalized query latency is actually somewhat questionable. Using multiple CPUs in parallel only decreases latency in the case of intra-query parallelism, not in the case of inter-query parallelism. Nonetheless, because we do not know exactly which type of parallelism a group employed in their runs, we decided to apply CPU normalization to both efficiency measures, throughput and latency.

Unfortunately, however, neither CPU normalization nor cost normalization are completely satisfying. The number of CPUs in a system does not say anything about the performance of these CPUs. The total cost of a computer, on the other hand, might include components like graphics cards and additional hard drives that were not used in a run at all.

Group	Comp. Run	Indri	Wumpus	Zettair
umass.allan	n/a	7.38/6.04/32-bit	0.48/0.32/32-bit	1.83/0.92/32-bit
max-planck.theobald	mpiiotopk2		0.23/0.18/64-bit	
rmit.scholer	rmit06effic	5.89/5.24/32-bit	0.39/0.30/32-bit	1.50/1.10/32-bit
lowlands-team.deVries	CWI06DISK1	4.64/4.17/64-bit	0.22/0.14/64-bit	0.93/0.81/32-bit
uwaterloo-clarke	uwmtFdcp12	4.49/3.02/64-bit	0.24/0.11/64-bit	1.62/0.62/32-bit

Table 5: *Comparative efficiency runs using Indri, Wumpus, and Zettair. Each field contains the average query latency in seconds (left), the average CPU time per query in seconds (middle), and the CPU type (right), either 32-bit or 64-bit.*

In order to obtain more reliable comparative performance results, we analyzed the efficiency numbers reported as part of the comparative efficiency runs conducted with Indri, Wumpus, and Zettair. Comparative runs were submitted by 4 out of 8 participating groups. In addition, Trevor Strohman from the University of Massachusetts (umass.allan) submitted three comparative runs in order to help us with our data sparseness problem. The results are shown in Table 5.

When examining the reported efficiency numbers, we noticed that, while performance figures were largely consistent between two different 32-bit systems (for both umass.allan and rmit.scholer, for instance, Wumpus is about 3.8 times faster than Zettair, while Indri is about 4 times slower than Zettair), the situation is different when comparing 32-bit hardware with 64-bit hardware. On uwaterloo-clarke's hardware, Wumpus is about 6.8 times faster than Zettair, while Indri is only 2.8 times slower than Zettair.

The large discrepancy between Wumpus and Zettair is due to Wumpus being designed for 64-bit and using 64-bit integer arithmetic throughout. When compiled for 32-bit, all 64-bit integer instructions have to simulated by sequences of 32-bit integer instructions, a transformation that is very costly. Zettair, on the other hand, uses 32-bit integer arithmetic and is compiled for a 32-bit architecture. Therefore, it does not experience the same slowdown when executed on a 32-bit computer. When uwaterloo-clarke recompiled Wumpus for 32-bit and ran it on their 64-bit hardware, this discrepancy almost vanished, and Wumpus was only 5 times faster than Zettair.

However, the non-proportional performance difference between the three retrieval systems when moving between different hardware configurations is not only because of CPU issues, but also because of different hard drive performance in the individual computer systems: uwaterloo-clarke stored index structures on a single disk, while umass.allan stored the index on a 3-disk RAID, lowlands-team.deVries on a 10-disk RAID, and rmit.scholer even on a 12-disk RAID. Because Indri, Wumpus, and Zettair produce inverted files of different size, the hard drive configuration has different effects on the three systems. For example, while Wumpus exhibits about the same performance on the hardware configurations used by lowlands-team.deVries and uwaterloo-clarke (220 vs. 240 ms per query), Zettair is 75% faster on lowlands-team.deVries's hardware than on uwaterloo-clarke's (930 vs. 1620 ms).

Unfortunately, the effect of different hard disk configurations cannot be eliminated by comparing CPU times, either. A system that uses software RAID, for instance, usually exhibits higher CPU utilization than a system using hardware RAID.

Despite these irregularities, we tried to come up with a tentative performance normalization procedure that allows us to compare the performance of different retrieval systems across hardware configuration boundaries. Because all four groups participating in the comparative efficiency task

Run	HW performance	Latency	Normalized latency	P@20	MRR
CWI06DISK1	CPU: 2.23, Total: 2.23	197 ms	439 ms ... 440 ms	0.4720	0.196
mpiiotopk2	CPU: 1.71, Total: 2.15	75 ms	128 ms ... 161 ms	0.4330	0.280
rmit06effic	CPU: 1.05, Total: 1.26	2,202 ms	2,304 ms ... 2,768 ms	0.4650	0.258
uwmtFdcp12	CPU: 2.89, Total: 2.01	32 ms	64 ms ... 92 ms	0.4790	0.219

Table 6: *Normalized efficiency based on the true performance of the underlying hardware configuration (estimated through Wumpus). Hardware performance is given relative to umass.allan (2.6 GHz Pentium IV with 2 GB RAM and a 3-way software RAID-0).*

submitted a comparative run using Wumpus, we chose to use Wumpus to establish the true performance of the underlying hardware and to normalized the latency numbers reported by the groups based on the performance estimate obtained through Wumpus. For each group, two performance estimates were obtained, one based on CPU time, the other based on average query latency. This led to a normalized efficiency interval instead of a single efficiency number. The results are shown in Table 6. Because of all the difficulties explained above, the outcome of the normalization process should be taken with a grain of salt.

If we have learned anything from our attempt to conduct a fair performance comparisons of different retrieval systems, then it is the insight that such a comparison is incredibly difficult, if not impossible, to achieve. The assumption that all document retrieval systems are relatively similar to each other and thus have similar performance characteristics (and by that we do not mean raw latency or throughput values) does not hold. It is entirely possible that moving to a different hardware configuration improves the performance of retrieval system A while depleting the performance of system B, as documented by Table 5

6 The Future of the Terabyte Track

2006 is the final year of the Terabyte Track in its current form. After three years, we have a reasonable number of topics and judgments, and we cannot see significant value in another year of similar experiments on the same collection.

In the future, we hope to resurrect the track with a substantially larger collection and an renewed focused on Web retrieval. Along with the standard adhoc and named page finding tasks, we plan to examine problems such as Web spam filtering and snippet extraction.

Acknowledgments

We thank everyone who helped to establish and operate the track over the past three years, particularly Yaniv Bernstein, Nick Craswell, David Hawking, Doug Oard, Falk Scholer and Ellen Voorhees.

References

[1] Yaniv Bernstein and Justin Zobel. A scalable system for identifying co-derivative documents. In *Proceedings of the Symposium on String Processing and Information Retrieval*, pages 55–67, Padova, Italy, 2004.

[2] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling. In *Proceedings of SIGIR 2006*, 2006.

[3] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, Sheffield, UK, 2004.

[4] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2004 Terabyte Track. In *Proceedings of the Thirteenth Text REtrieval Conference*, Gaithersburg, MD, November 2004. NIST Special Publication 500-261. See `trec.nist.gov`.

[5] C. L. A. Clarke, F. Scholer, and I. Soboroff. The TREC 2005 Terabyte Track. In *Proceedings of the Fourteenth Text REtrieval Conference*, Gaithersburg, MD, November 2005. NIST Special Publication 500-266. See `trec.nist.gov`.

[6] Ellen M. Voorhees. Overview of the TREC 2005 robust retrieval track. In *Proceedings of the Fourteenth Text REtrieval Conference*, Gaithersburg, MD, November 2005. NIST Special Publication 500-266. See `trec.nist.gov`.

[7] Emine Yilmaz and Javed Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of CIKM 2006, to appear*, 2006.

Text REtrieval Conference (TREC)
to encourage research in information retrieval
from large text collections
SP 500-272
NIST
TREC 2006
Conference
Proceedings

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 the 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.

National Construction Safety Team Act Reports—This series comprises the reports of investigations carried out under Public Law 107-231, the technical cause(s) of the building failure investigated; any technical recommendations for changes to or the establishment of evacuation and emergency response procedures; any recommended specific improvements to building standards, codes, and practices; and recomendations for research and other approprate actions to help prevent future building failures.

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.

U.S. Department of Commerce

National Institute of Standards
and Technology
Gaithersburg, MD 20899-0001