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

NIST Special Publication 500-266

Information Technology:

The Fourteenth Text Retrieval Conference

TREC 2005

Ellen M. Voorhees
and
Lori P. Buckland,
Editors

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

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

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

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

Office of the Director
- National Quality Program
- International and Academic Affairs

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

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

Manufacturing Extension Partnership Program
- Regional Programs
- National Programs
- Program Development

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

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

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

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

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

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

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

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

NIST Special Publication 500-266

Information Technology:

The Fourteenth Text Retrieval Conference

TREC 2005

Ellen M. Voorhees
and
Lori P. Buckland,
Editors

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

October 2006

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

Technology Administration
Robert Cresanti, Under Secretary of Commerce for Technology

National Institute of Standards and Technology
William Jeffrey, Director

Reports on Information Technology

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

National Institute of Standards and Technology Special Publication 500-266
Natl. Inst. Stand. Technol. Spec. Publ. 500-266, 162 pages (October 2006)

Foreword

This report constitutes the proceedings of the 2005 edition of the Text REtrieval Conference, TREC 2005, held in Gaithersburg, Maryland, November 15–18, 2005. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Advanced Research and Development Activity (ARDA). Approximately 200 people attended the conference, including representatives from 23 different countries. The conference was the fourteenth 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.

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

Ellen Voorhees
September 20, 2006

TREC 2005 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
John Prange, U.S. Department of Defense
Steve Robertson, Microsoft
Mark Sanderson, University of Sheffield
Ian Soboroff, NIST
Karen Sparck Jones, University of Cambridge
Ross Wilkinson, CSIRO

TREC 2005 Proceedings

Overview Papers

Other Papers

(contents of these papers are found on the TREC 2005 Proceedings CD)

Appendix

(contents of the Appendix are found on the TREC 2005 Proceedings CD)

Common Evaluation Measures

Enterprise Discussion Runs
Enterprise Discussion Results
Enterprise Expert Runs
Enterprise Expert Results
Enterprise Known-Item Runs
Enterprise Known-Item Results

Genomics adhoc Runs
Genomics adhoc Results
Genomics Category Runs
Genomics Category Results

HARD Baseline Runs
HARD Baseline Results

HARD Final Runs
HARD Final Results

QA Document-Ranking Runs
QA Document-Ranking Results
QA Main Runs
QA Main Results
QA Relationship Runs
QA Relationship Results

Robust Runs
Robust Results

Spam Runs
Spam Results

Terabyte adhoc Runs
Terabyte adhoc Results
Terabyte Efficiency Runs
Terabyte Efficiency Results
Terabyte Named-Page Runs
Terabyte Named-Page Results

Papers: Alphabetical by Organization

(contents of these papers are found on the TREC 2005 Proceedings CD)

Chinese Academy of Sciences
NLPR at TREC 2005: HARD Experiments

CAS-ICT at TREC 2005 Robust Track: Using Query Expansion and RankFusion to Improve Effectiveness and Robustness of Ad Hoc Information Retrieval

CAS-ICT at TREC 2005 SPAM Track: Using Non-Textual Information to Improve Spam Filtering Performance

NLPR at TREC 2005: HARD Experiments

Relevance Feedback by Exploring the Different Feedback Sources and Collection Structure

Chinese University of Hong Kong
Pattern-Based Customized Learning for TREC Genomics Track Categorization Task

MATRIX at the TREC 2005 Robust Track

City University of New York
MATRIX at the TREC 2005 Robust Track

CL Research
Exploring Document Content with XML to Answer Questions

Columbia University

CSIRO ICT Centre
TREC 14 Enterprise Track at CSIRO and ANU

CWI
Overview of the TREC 2005 Enterprise Track

The Lowlands' TREC Experiments 2005

Dalhousie University
DalTREC 2005 QA System Jellyfish:
Mark-and-Match Approach to Question Answering

DalTREC 2005 Spam Track: Spam Filtering Using N-gram-based Techniques

Dalian University of Technology
TREC 2005 Genomics Track Experiments at DUTAI

OOO Datapark
DataparkSearch at TREC 2005

Dragon Development Corporation
QACTIS-based Question Answering at TREC 2005

Drexel University
TREC 2005 Enterprise Track Results from Drexel

Dublin City University
Structural Term Extraction for Expansion of Template-Based Genomic Queries

Dublin City University at the TREC 2005 Terabyte Track

Ecole Nationale Superieure des Mines de Sanit-Etienne
Fuzzy Proximity Ranking with Boolean Queries

Embratel
CRM114 versus Mr. X: CRM114 Notes for the TREC 2005 Spam Track

Erasmus University Medical Center
TREC 2005 Genomics Track A Concept-Based Approach to Text Categorization

Freie Universitat Berlin
CRM114 versus Mr. X: CRM114 Notes for the TREC 2005 Spam Track

Fudan University
WIM at TREC 2005

FDUQA on TREC 2005 QA Track

Google, Inc.
IBM's PIQUANT II in TREC 2005

Harbin Institute of Technology
Insun05QA on QA Track of TREC 2005

Henggeler Computer Consultants
QACTIS-based Question Answering at TREC 2005

The Hong Kong Polytechnic University
MATRIX at the TREC 2005 Robust Track

Hummingbird
Enterprise, QA, Robust and Terabyte Experiments with Hummingbird SearchServer at TREC 2005

IBM Almaden Research Center
Thresholding Strategies for Text Classifiers: TREC 2005 Biomedical Triage Task Experiments

IBM Haifa Research Lab
Juru at TREC 2005: Query Prediction in the Terabyte and the Robust Tracks

IBM India Research Lab
Biomedical Document Triage: Automatic Classification Exploiting Category Specific Knowledge

IBM Research
IBM SpamGuru on the TREC 2005 Spam Track

IBM T.J. Watson Research
TREC 2005 Genomics Track Experiments at IBM Watson

IBM's PIQUANT II in TREC 2005

Indian Institute of Technology
Biomedical Document Triage: Automatic Classification Exploiting Category Specific Knowledge

Illinois Institute of Technology
IIT TREC 2005: Genomics Track

Indiana University, Bloomington
WIDIT in TREC 2005 HARD, Robust, and SPAM Tracks

Institute for Infocomm Research
TREC 2005 Genomics Track at I2R

IRIT
A Conceptual Indexing Approach for the TREC Robust Task

Johns Hopkins Applied Physics Laboratory
QACTIS-based Question Answering at TREC 2005 JHU/APL at TREC 2005: QA Retrieval and Robust Tracks

Jozef Stefan Institute
Spam Filtering Using Character-Level Markov Models: Experiments for the TREC 2005 Spam Track

Language Computer Corporation
Employing Two Question Answering Systems in TREC 2005

LexiClone
Differential Linguistics at NIST TREC

The Lowlands' Team
The Lowlands' TREC Experiments 2005

Macquarie University
AnswerFinder at TREC 2005

Massey University
A TREC Along the Spam Track with SpamBayes

Meiji University
Meiji University HARD and Robust Track Experiments

Microsoft Research
Factoid Question Answering over Unstructured and Structured Web Content

Microsoft Research Asia
Research on Expert Search at Enterprise
Track of TREC 2005

Microsoft Research Cambridge
Microsoft Cambridge at TREC 14: Enterprise Track

Overview of the TREC 2005 Enterprise Track

MIT
External Knowledge Sources for Question Answering

The MITRE Corporation
MITRE's Qanda at TREC 14

Mitsubishi
CRM114 versus Mr. X: CRM114 Notes for the TREC 2005 Spam Track

Nankai University
Research on Expert Search at Enterprise Track of TREC 2005

National Institute of Standards and Technology (NIST)
Overview of TREC 2005

Overview of the TREC 2005 Enterprise Track

Overview of the TREC 2005 Question Answering Track

Overview of the TREC 2005 Robust Retrieval Track

The TREC 2005 Terabyte Track

National Library of Medicine
Fusion of Knowledge-Intensive and Statistical Approaches for Retrieving and Annotating Textual Genomics Documents

National Taiwan University
Retrieval of Biomedical Documents by Prioritizing Key Phrases

Identifying Relevant Full-Text Articles for Database Curation

National Taiwan University at Terabyte Track of TREC 2005

National University of Singapore
Using Syntactic and Semantic Relation Analysis in Question Answering

Oregon Health & Science University
TREC 2005 Genomics Track Overview

A Comparison of Techniques for Classification and Ad Hoc Retrieval of Biomedical Documents

Peking University
CNDS Expert Finding System for TREC 2005

Peking University at the TREC 2005 Question and Answering Track

Tianwang Search Engine at TREC 2005: Terabyte Track

PUC/RS
Simple Language Models for Spam Detection

Queen Mary University of London
The QMUL Team with Probabilistic SQL at Enterprise Track

Queens College, CUNY
TREC 2005 Robust Track Experiments Using PIRCS

Queensland University of Technology
Queensland University of Technology at TREC 2005

Queen's University
Applying Probabilistic Thematic Clustering for Classification in the TREC 2005 Genomics Track

RMIT University
The TREC 2005 Terabyte Track

RMIT University at TREC 2005: Terabyte and Robust Track

Rutgers University
DIMACS at the TREC 2005 Genomics Track

Rutgers Information Interaction Lab at TREC 2005: Trying HARD

Sabir Research, Inc.
Looking at Limits and Tradeoffs: Sabir Research at TREC 2005

Saarland University
QuALiM at TREC 2005: Web-Question Answering with FrameNet

TREC 2005 Question Answering Experiments at Tokyo Institute of Technology

Science Applications International Corporation (SAIC)
SAIC & University of Virginia at TREC 2005: HARD Track

Shanghai Jiaotong University
Research on Expert Search at Enterprise Track of TREC 2005

Simon Fraser University
Synonym-Based Expansion and Boosting-Based Re-Ranking: A Two-phase Approach for Genomic Information Retrieval

SIM University
Report on the TREC 2005 Experiment: Genomics Track

Stanford University
Classifying Biomedical Articles by Making Localized Decisions

State University of New York at Buffalo
Experiments on Genomics Ad Hoc Retrieval

State University of New York at Stony Brook
Question Answering with Lydia (TREC 2005 QA Track)

Temple University
TREC 2005 Enterprise Track Results from Drexel

TNO
TREC 2005 Genomics Track A Concept-Based Approach to Text Categorization

Tokyo Institute of Technology
TREC 2005 Question Answering Experiments at Tokyo Institute of Technology

Tsinghua University (State Key Lab)
Learning Domain-Specific Knowledge from Context--THUIR at TREC 2005 Genomics Track

THUIR at TREC 2005: Enterprise Track

THUIR at TREC 2005 Terabyte Track

U.S. Department of Defense
QACTIS-based Question Answering at TREC 2005

University of Alaska, Arctic Region Supercomputing Center
Logistic Regression Merging of Amberfish and Lucene Multisearch Results

University of Albany, SUNY
ILQUA--An IE-Driven Question Answering System

University of Amsterdam
Combining Thesauri-Based Methods for Biomedical Retrieval

Effective Smoothing for a Terabyte of Text

Language Modeling Approaches for Enterprise Tasks

Towards a Multi-Stream Question Answering-As-XML-Retrieval Strategy

University of California
CRM114 versus Mr. X: CRM114 Notes for the TREC 2005 Spam Track

University of California, Berkeley
TREC 2005 Genomics Track Overview

University of Central Florida
Question Answering with SEMEX at TREC 2005

University of Colorado Health Sciences Center
Concept Recognition and the TREC Genomics Tasks

University of Duisburg-Essen
Applying the Annotation View on Messages for Discussion Search

University of Edinburgh
Question Answering with QED at TREC 2005

University of Essex
Question Answering Using the DLT System at TREC 2005

University of Glasgow
University of Glasgow at TREC 2005: Experiments in Terabyte and Enterprise Tracks with Terrier

University Hospital of Geneva
Report on the TREC 2005 Experiment: Genomics Track

Evaluation of Stemming, Query Expansion and Manual Indexing Approaches for the Genomic Task

University of Illinois at Chicago
Experiment Report of TREC 2005 Genomics Track ad hoc Retrieval Task

UIC at TREC 2005: Robust Track

University of Illinois at Urbana-Champaign
An Axiomatic Approach to IR--UIUC TREC 2005 Robust Track Experiments

Interactive Construction of Query Language Models - UIUC TREC 2005 HARD Track Experiments

UIUC/MUSC at TREC 2005 Genomics Track

The University of Iowa
Experiments in Questions and Relationships at the University of Iowa

University of Limerick
Question Answering Using the DLT System at TREC 2005

University of Maryland, Baltimore County
Fusion of Knowledge-Intensive and Statistical Approaches for Retrieving and Annotating Textual Genomics Documents

University of Maryland, College Park
Fusion of Knowledge-Intensive and Statistical Approaches for Retrieving and Annotating Textual Genomics Documents

A Menagerie of Tracks at Maryland: HARD, Enterprise, QA, and Genomics, Oh My!

University of Massachusetts, Amherst
Hard Track Overview in TREC 2005 High Accuracy Retrieval from Documents

UMass Robust 2005: Using Mixtures of Relevance Models for Query Expansion

Indri at TREC 2005: Terabyte Track

The University of Melbourne
Melbourne University 2005: Enterprise and Terabyte Tasks

RMIT University at TREC 2005: Terabyte and Robust Track

The University of Michigan-Dearborn
UM-D at TREC 2005: Genomics Track

Universite de Neuchatel
Report on the TREC 2005 Experiment: Genomics Track

Evaluation of Stemming, Query Expansion and Manual Indexing Approaches for the Genomic Task

University of North Carolina
University of North Carolina's HARD Track Experiment at TREC 2005

University of North Texas
UNT 2005 TREC QA Participation: Using Lemur as IR Search Engine

Universita degli Studi di Milano
MG4J at TREC 2005

University of Padova
Symbol-Based Query Expansion Experiments at TREC 2005 Genomics Track

University of Pennsylvania
TREC 2005 Genomics Track Experiments at IBM Watson

Universita di Pisa
IXE at the TREC 2005 Terabyte Task

University of Pittsburgh
Pitt at TREC 2005: HARD and Enterprise

Universitat Politecnica de Catalunya
TALP-UPC at TREC 2005: Experiments Using a Voting Scheme Among Three Heterogeneous QA Systems

University of Sheffield
The University of Sheffield's TREC 2005 Q&A Experiments

University of Strathclyde
University of Strathclyde at TREC HARD

Papers: Organized by Track

(contents of these papers are found on the TREC 2005 Proceedings CD)

Enterprise

Australian National University
TREC 14 Enterprise Track at CSIRO and ANU

Beijing University of Posts and Telecommunications
TREC 2005 Enterprise Track Experiments at BUPT

Bilkent University
Experiments for HARD and Enterprise Tracks

California State University San Marcos
CSUSM at TREC 2005: Genomics and Enterprise Track

Carnegie Mellon University
Experiments with Language Models for Known-Item Finding of E-mail Messages

CSIRO ICT Centre
TREC 14 Enterprise Track at CSIRO and ANU

The Lowlands' TREC Experiments 2005

CWI
Overview of the TREC 2005 Enterprise Track

Drexel University
TREC 2005 Enterprise Track Results from Drexel

Fudan University
WIM at TREC 2005

Microsoft Research Asia
Research on Expert Search at Enterprise Track of TREC 2005

Microsoft Research Ltd.
Microsoft Cambridge at TREC 14: Enterprise Track

Microsoft Research Cambridge
Overview of the TREC 2005 Enterprise Track

Nankai University
Research on Expert Search at Enterprise Track of TREC 2005

NIST
Overview of the TREC 2005 Enterprise Track

Peking University
CNDS Expert Finding System for TREC 2005

Queen Mary University of London
The QMUL Team with Probabilistic SQL at Enterprise Track

Shanghai Jiaotong University
Research on Expert Search at Enterprise Track of TREC 2005

Temple University
TREC 2005 Enterprise Track Results from Drexel

Tsinghua University (State Key Lab)
THUIR at TREC 2005: Enterprise Track

University of Amsterdam
Language Modeling Approaches for Enterprise Tasks

University of Duisburg-Essen
Applying the Annotation View on Messages for Discussion Search

University of Glasgow
University of Glasgow at TREC 2005: Experiments in Terabyte and Enterprise Tracks with Terrier

University of Maryland College Park
A Menagerie of Tracks at Maryland: HARD, Enterprise, QA, and Genomics, Oh My!

The University of Melbourne
Melbourne University 2005: Enterprise and Terabyte Tasks

University of Pittsburgh
Pitt at TREC 2005: HARD and Enterprise

University of Twente
The Lowlands' TREC Experiments 2005

University of Waterloo
Experiments for HARD and Enterprise Tracks

Genomics

Academia Sinica
Enhance Genomic IR with Term Variation and Expansion: Experience of the IASL Group at Genomic Track 2005

Arizona State University
Genomic Information Retrieval Through Selective Extraction and Tagging by the ASU-BioAL Group

California State University San Marcos
CSUSM at TREC 2005: Genomics and Enterprise Track

Carnegie Mellon University
Thresholding Strategies for Text Classifiers: TREC 2005 Biomedical Triage Task Experiments

Chinese University of Hong Kong
Pattern-Based Customized Learning for TREC Genomics Track Categorization Task

OOO Datapark
DataparkSearch at TREC 2005

Dalian University of Technology
TREC 2005 Genomics Track Experiments at DUTAI

Dublin City University
Structural Term Extraction for Expansion of Template-Based Genomic Queries

Erasmus University Medical Center
Notebook Paper TREC 2005 Genomics Track

Fudan University
WIM at TREC 2005

Harbin Institute of Technology
Insun05QA on QA Track of TREC 2005

IBM India Research Lab
Biomedical Document Triage: Automatic Classification Exploiting Category Specific Knowledge

IBM Watson Research
TREC 2005 Genomics Track Experiments at IBM Watson

Idec Corporation
TREC 2005 Genomics Overview

Illinois Institute of Technology
IIT TREC 2005: Genomics Track

Indian Institute of Technology
Biomedical Document Triage: Automatic Classification Exploiting Category Specific Knowledge

Institute for Infocomm Research
TREC 2005 Genomics Track at I2R

Medical University of South Carolina
UIUC/MUSC at TREC 2005 Genomics Track

National Library of Medicine
Fusion of Knowledge-Intensive and Statistical Approaches for Retrieving and Annotating Textual Genomics Documents

National Taiwan University
Retrieval of Biomedical Documents by Prioritizing Key Phrases

Identifying Relevant Full-Text Articles for Database Curation

A Comparison of Techniques for Classification and Ad Hoc Retrieval of Biomedical Documents

Oregon Health & Science University
TREC 2005 Genomics Track Overview

Queen's University
Applying Probabilistic Thematic Clustering for Classification in the TREC 2005 Genomics Track

Rutgers University, DIMACS
DIMACS at the TREC 2005 Genomics Track

Simon Fraser University
Synonym-Based Expansion and Boosting-Based Re-Ranking: A Two-phase Approach for Genomic Information Retrieval

State University of New York at Buffalo
Experiments on Genomics Ad Hoc Retrieval

TNO
Notebook Paper TREC 2005 Genomics Track

Tsinghua University (State Key Lab)
Learning Domain-Specific Knowledge from Context--THUIR at TREC 2005 Genomics Track

SIM University and University Hospital of Geneva
Report on the TREC 2005 Experiment: Genomics Track

Stanford University
Classifying Biomedical Articles by Making Localized Decisions

University of Amsterdam
Combining Thesauri-Based Methods for Biomedical Retrieval

University of California, Berkeley
TREC 2005 Genomics Track Overview

University of Colorado Health Sciences Center
Concept Recognition and the TREC Genomics Tasks

University Hospital of Geneva
Evaluation of Stemming, Query Expansion and Manual Indexing Approaches for the Genomic Task

University of Illinois at Chicago
Experiment Report of TREC 2005 Genomics Track ad hoc Retrieval Task

University of Illinois at Urbana-Champaign
UIUC/MUSC at TREC 2005 Genomics Track

The University of Iowa
Experiments in Questions and Relationships at the University of Iowa

University of Maryland, Baltimore County
Fusion of Knowledge-Intensive and Statistical Approaches for Retrieving and Annotating Textual Genomics Documents

University of Maryland, College Park
Fusion of Knowledge-Intensive and Statistical Approaches for Retrieving and Annotating Textual Genomics Documents

A Menagerie of Tracks at Maryland: HARD, Enterprise, QA, and Genomics, Oh My!

The University of Michigan-Dearborn
UM-D at TREC 2005: Genomics Track

University of Neuchatel
Report on the TREC 2005 Experiment: Genomics Track

Evaluation of Stemming, Query Expansion and Manual Indexing Approaches for the Genomic Task

University of Padova
Symbol-Based Query Expansion Experiments at TREC 2005 Genomics Track

University of Pennsylvania
TREC 2005 Genomics Track Experiments at IBM Watson

University of Tampere
TREC 2005 Genomics Track Experiments at UTA

Yahoo, Inc.
TREC 2005 Genomics Track Experiments at IBM Watson

University of Wisconsin, Madison
Classifying Biomedical Articles by Making Localized Decisions

York University
York University at TREC 2005: Genomics Track

Hard

Bilkent University
Experiments for HARD and Enterprise Tracks

Chinese Academy of Sciences
NLPR at TREC 2005: HARD Experiments

Relevance Feedback by Exploring the Different Feedback Sources and Collection Structure

CWI
The Lowlands' TREC Experiments 2005

Indiana University, Bloomington
WIDIT in TREC 2005 HARD, Robust, and SPAM Tracks

Meiji University
Meiji University HARD and Robust Track Experiments

Rutgers University
Rutgers Information Interaction Lab at TREC 2005: Trying HARD

Science Applications International Corporation (SAIC)
SAIC & University of Virginia at TREC 2005: HARD Track

University of Illinois at Urbana-Champaign
Interactive Construction of Query Language Models--UIUC TREC 2005 HARD Track Experiments

University of Maryland, College Park
A Menagerie of Tracks at Maryland: HARD, Enterprise, QA, and Genomics, Oh My!

University of Massachusetts, Amherst
HARD Track Overview in TREC 2005 High Accuracy Retrieval from Documents

When Less is More: Relevance Feedback Falls Short and Term Expansion Succeeds at HARD 2005

University of North Carolina
University of North Carolina's HARD Track Experiment at TREC 2005

University of Pittsburgh
Pitt at TREC 2005: HARD and Enterprise

University of Strathclyde
University of Strathclyde at TREC HARD

University of Twente
The Lowlands' TREC Experiments 2005

University of Virginia
SAIC & University of Virginia at TREC 2005:
HARD Track

University of Waterloo
Experiments for HARD and Enterprise Tracks

York University
York University at TREC 2005: HARD Track

QA

Arizona State University
Building on Redundancy: Factoid Question Answering, Robust Retrieval and the "Other"

Carnegie Mellon University
JAVELIN I and II Systems at TREC 2005

CL Research
Exploring Document Content with XML to Answer Questions

Dalhousie University
DalTREC 2005 QA System Jellyfish: Mark-and-Match Approach to Question Answering

Dragon Development Corporation
QACTIS-based Question Answering at TREC 2005

Fudan University
FDUQA on TREC 2005 QA Track

Henggeler Computer Consultants
QACTIS-based Question Answering at TREC 2005

Hummingbird
Enterprise, QA, Robust and Terabyte Experiments with Hummingbird SearchServer at TREC 2005

IBM T.J. Watson Research Center
IBM's PIQUANT II in TREC 2005

Johns Hopkins Applied Physics Laboratory
JHU/APL at TREC 2005: QA Retrieval and Robust Tracks

QACTIS-based Question Answering at TREC 2005

Langugage Computer Corporation
Employing Two Question Answering Systems in TREC 2005

LexiClone
Differential Linguistics at NIST TREC

Macquarie University
AnswerFinder at TREC 2005

Microsoft Research
Factoid Question Answering over Unstructured and Structured Web Content

MIT Computer Science and Artificial Intelligence Laboratory
External Knowledge Sources for Question Answering

The MITRE Corporation
MITRE's Qanda at TREC 14

National Institute of Standards and Technology
Overview of the TREC 2005 Question Answering Track

National University of Singapore
Using Syntactic and Semantic Relation Analysis in Question Answering

Peking University
Peking University at the TREC 2005 Question and Answering Track

Saarland University
TREC 2005 Question Answering Experiments at Tokyo Institute of Technology

Saarland University/DFKI GmbH
QuALiM at TREC 2005: Web-Question Answering with FrameNet

Sabir Research, Inc.
Looking at Limits and Tradeoffs: Sabir Research at TREC 2005

State University of New York at Stony Brook
Question Answering with Lydia (TREC 2005 QA Track)

Tokyo Institute of Technology
TREC 2005 Question Answering Experiments at Tokyo Institute of Technology

University of Albany SUNY
ILQUA -- An IE-Driven Question Answering System

University of Amsterdam
Towards a Multi-Stream Question Answering-As-XML-Retrieval Strategy

University of Central Florida
Question Answering with SEMEX at TREC 2005

University of Edinburgh
Question Answering with QED at TREC 2005

University of Essex
Question Answering Using the DLT System at TREC 2005

University of Limerick
Question Answering Using the DLT System at TREC 2005

University of Maryland, College Park
A Menagerie of Tracks at Maryland: HARD, Enterprise, QA, and Genomics, Oh My!

University of North Texas
UNT 2005 TREC QA Participation: Using Lemur as IR Search Engine

Universitat Politecnica de Catalunya
TALP-UPC at TREC 2005: Experiments Using a Voting Scheme Among Three Heterogeneous QA Systems

University of Sheffield
The University of Sheffield's TREC 2005 Q&A Experiments

University of Sydney
Question Answering with QED at TREC 2005

U.S. Department of Defense
QACTIS-based Question Answering at TREC 2005

Robust

Arizona State University
Building on Redundancy: Factoid Question Answering, Robust Retrieval and the "Other"

Columbia University
Building on Redundancy: Factoid Question Answering, Robust Retrieval and the "Other"

Chinese Academy of Sciences
CAS-ICT at TREC 2005 Robust Track: Using Query Expansion and RankRusion to Improve Effectiveness and Robustness of Ad Hoc Information Retrieval

The Chinese University of Hong Kong
MATRIX at the TREC 2005 Robust Track

City University of New York
MATRIX at the TREC 2005 Robust Track

Ecole Nationale Superieure des Mines de Sanit-Etienne
Fuzzy Proximity Ranking with Boolean Queries

Hummingbird
Enterprise, QA, Robust and Terabyte Experiments with Hummingbird SearchServer at TREC 2005

IBM Haifa Research Lab
Juru at TREC 2005: Query Prediction in the Terabyte and the Robust Tracks

Indiana University, Bloomington
WIDIT in TREC 2005 HARD, Robust, and SPAM Tracks

IRIT-CNRS
A Conceptual Indexing Approach for the TREC Robust Task

IRIT-UPS
A Conceptual Indexing Approach for the TREC Robust Task

Johns Hopkins Applied Physics Laboratory
JHU/APL at TREC 2005: QA Retrieval and Robust Tracks

Meiji University
Meiji University HARD and Robust Track Experiments

National Institute of Standards and Technology
Overview of the TREC 2005 Robust Retrieval Track

Queens College, CUNY
TREC 2005 Robust Track Experiments Using PIRCS

Queensland University of Technology
Queensland University of Technology at TREC 2005

RMIT University
RMIT University at TREC 2005: Terabyte and Robust Track

Sabir Research, Inc.
Looking at Limits and Tradeoffs: Sabir Research at TREC 2005

The University of Hong Kong
Building on Redundancy: Factoid Question Answering, Robust Retrieval and the "Other"

University of Illinois at Chicago
UIC at TREC 2005: Robust Track

University of Illinois at Urbana-Champaign
An Axiomatic Approach to IR--UIUC TREC 2005 Robust Track Experiments

University of Massachusetts, Amherst
UMass Robust 2005: Using Mixtures of Relevance Models for Query Expansion

The University of Melbourne
RMIT University at TREC 2005: Terabyte and Robust Track

Spam

Beijing University of Posts and Telecommunications
PRIS Kidult Anti-SPAM Solution at the TREC 2005 Spam Track: Improving the Performance of Naive Bayes for Spam Detection

Breyer, L.A.
DBACL at the TREC 2005

Chinese Academy of Sciences (NLPR)
CAS-ICT at TREC 2005 SPAM Track: Using Non-Textual Information to Improve Spam Filtering Performance

Dalhousie University
DalTREC 2005 Spam Track: Spam Filtering Using N-gram-based Techniques

Embratel
CRM114 versus Mr. X: CRM114 Notes for the TREC 2005 Spam Track

Freie Universitat Berlin
CRM114 versus Mr. X: CRM114 Notes for the TREC 2005 Spam Track

IBM Research
IBM SpamGuru on the TREC 2005 Spam Track

Indiana University, Bloomington
WIDIT in TREC 2005 HARD, Robust, and SPAM Tracks

Josef Stefan Institute
Spam Filtering Using Character-Level Markov Models: Experiments for the TREC 2005 Spam Track

L.A. Breyer
TREC 2005 Enterprise Track Experiments at BUPT

Massey University
A TREC Along the Spam Track with SpamBayes

Mitsubishi
CRM114 versus Mr. X: CRM114 Notes for the TREC 2005 Spam Track

Mitsubishi and University of California
CRM114 versus Mr. X: CRM114 Notes for the TREC 2005 Spam Track

PUC/RS
Simple Language Models for Spam Detection

University of Waterloo
TREC 2005 Spam Track Overview

York University
York University at TREC 2005: SPAM Track

Terabyte

Dublin City University
Dublin City University at the TREC 2005 Terabyte Track

Hummingbird
Enterprise, QA, Robust and Terabyte Experiments with Hummingbird SearchServer at TREC 2005

IBM Haifa Research Lab
Juru at TREC 2005: Query Prediction in the Terabyte and the Robust Tracks

National Institute of Standards and Technology
The TREC 2005 Terabyte Track

National Taiwan University
National Taiwan University at Terabyte Track of TREC 2005

Peking University
Tianwang Search Engine at TREC 2005: Terabyte Track

Queensland University of Technology
Queensland University of Technology at TREC 2005

RMIT
The TREC 2005 Terabyte Track

RMIT University at TREC 2005: Terabyte and Robust Track

Sabir Research, Inc.
Looking at Limits and Tradeoffs: Sabir Research at TREC 2005

Tsinghua University (State Key Lab)
THUIR at TREC 2005 Terabyte Track

University of Amsterdam
Effective Smoothing for a Terabyte of Text

Universita degli Studi di Milano
MG4J at TREC 2005

University of Glasgow
University of Glasgow at TREC 2005: Experiments in Terabyte and Enterprise Tracks with Terrier

University of Massachusetts, Amherst
Indri at TREC 2005: Terabyte Track

The University of Melbourne
RMIT University at TREC 2005: Terabyte and Robust Track

Melbourne University 2005: Enterprise and Terabyte Tasks

Universita di Pisa
IXE at the TREC 2005 Terabyte Task Efficiency vs. Effectiveness in Terabyte-Scale Information Retrieval

University of Waterloo
The TREC 2005 Terabyte Track

York University
York University at TREC 2005: Terabyte Track

Abstract

This report constitutes the proceedings of the 2005 edition of the Text REtrieval Conference, TREC 2005, held in Gaithersburg, Maryland, November 15–18, 2005. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Advanced Research and Development Activity (ARDA). TREC 2005 had 117 participating groups including participants from 23 different countries.

TREC 2005 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: detecting spam in an email stream, enterprise search, question answering, retrieval in the genomics domain, improving the consistency of retrieval systems across queries, improving retrieval effectiveness by focusing on user context, and retrieval from terabyte-scale collections.

The conference included paper sessions and discussion groups. The overview papers for the different "tracks" and for the conference as a whole are gathered in this bound version of the proceedings. The papers from the individual participants and the evaluation output for the runs submitted to TREC 2005 are contained on the disk included in the volume. The TREC 2005 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 2005

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

1 Introduction

The fourteenth Text REtrieval Conference, TREC 2005, was held at the National Institute of Standards and Technology (NIST) 15 to 18 November 2005. The conference was co-sponsored by NIST and the US Department of Defense Advanced Research and Development Activity (ARDA). TREC 2005 had 117 participating groups from 23 different countries. Table 2 at the end of the paper lists the participating groups.

TREC 2005 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 2005 contained seven areas of focus called "tracks". Two tracks focused on improving basic retrieval effectiveness by either providing more context or by trying to reduce the number of queries that fail. Other tracks 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 specific tasks performed in each of the tracks are summarized in Section 3 below.

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 MEDLINE record, a web page, or an email message.

The prototypical retrieval task is a researcher doing a literature search in a library. In this environment the retrieval system knows the set of documents to be searched (the library's holdings), but cannot anticipate the particular topic that will be investigated. We call this an *ad hoc* retrieval task, reflecting the arbitrary subject of the search and its short duration. Other examples of ad hoc searches are web surfers using Internet search engines, lawyers performing patent searches or looking for precedences in case law, and analysts searching archived news reports for particular events. A retrieval system's response to an ad hoc search is generally a list of documents ranked by decreasing similarity to the query. Most of the TREC 2005 tracks included some sort of an ad hoc search task.

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 and the known-item task within the enterprise track are examples of known-item search tasks.

In a *categorization* task, the system is responsible for assigning a document to one or more categories from among a given set of categories. In the spam track, deciding whether a given mail message is spam is a categorization task; the genomics track had several categorization tasks in TREC 2005 as well.

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 expert-finding task in the enterprise track is a type 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 [2, 6], and TREC continues this tradition. A test collection is an abstraction of an operational retrieval environment that provides a means for researchers to explore the relative benefits of different retrieval strategies in a laboratory setting. Test collections consist of three parts: a set of documents, a set of information needs (called *topics* in TREC), and *relevance judgments*, an indication of which documents should be retrieved in response to which topics. 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 primary 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

```
<num> Number:  758
<title> Embryonic stem cells
<desc> Description:  What are embryonic stem cells, and what
restrictions are placed on their use in research?
<narr> Narrative:  Explanation of the nature of embryonic stem cells is
relevant.  Their usefulness in research is relevant.  Sources for them
and restrictions on them also are relevant.
```

Figure 1: A sample TREC 2005 topic from the terabyte track test set.

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.

The primary TREC document sets consist mostly of newspaper or newswire articles. 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, the text was kept as close to the original as possible. No attempt was made to correct spelling errors, sentence fragments, strange formatting around tables, or similar faults.

2.1.2 Topics

TREC distinguishes between a statement of information need (the topic) and the data structure that is actually given to a retrieval system (the query). The TREC test collections provide topics to allow a wide range of query construction methods to be tested and also to include a clear statement of what criteria make a document relevant. The format of a topic statement has evolved since the earliest TRECs, but it has been stable since TREC-5 (1996). A topic statement generally consists of four sections: an identifier, a title, a description, and a narrative. An example topic taken from this year's 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 [4]. 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 [7].

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 [5] 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 [10]. 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 [9]. 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 [3].

The uniquely-retrieved-relevant-documents 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 this year. The retrieval test collection built in the TREC 2005 HARD and robust tracks has a demonstrable bias toward documents that contain topic title words. That is, a very large fraction of the known relevant documents for that collection contain many topic title words despite the fact that documents with fewer topic title words that would have been judged relevant exist in the collection. (Details are given in the robust track overview paper later in this volume [8].)

The bias results from pools that are shallow *relative to the number of documents in the collection.* Many otherwise diverse retrieval methodologies sensibly 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. But a large document set will contain many documents that include topic title words. 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. The robust track contained one run that did not concentrate on topic title words and could thus demonstrate the bias in the other runs. No such "smoking-gun" run exists for the collections built in the TREC 2004 and 2005 terabyte track, but a similar bias must surely exist in these collections. The biased collections are still useful for comparing retrieval methodologies that have a matching bias (and the results of the 2005 tracks are valid since the runs were used to build the collections), but results on these collections need to be interpreted judiciously when comparing methodologies that do not emphasize topic title words.

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, new evaluation measures have had to be devised. Indeed, developing an appropriate evaluation methodology for a new task is one of the primary goals of the TREC tracks. The details of the evaluation methodology used in a track are described in the track's overview paper.

3 TREC 2005 Tracks

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

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

This section describes the tasks performed in the TREC 2005 tracks. See the track reports later in these proceedings for a more complete description of each track.

3.1 The enterprise track

TREC 2005 was the first year for the enterprise track, which is an outgrowth of previous years' web track tasks. 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

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

Track	TREC													
	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
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
Web	—	—	—	—	—	—	—	17	23	30	23	27	18	—
Video	—	—	—	—	—	—	—	—	—	12	19	—	—	—
Novelty	—	—	—	—	—	—	—	—	—	—	13	14	14	—
Genomics	—	—	—	—	—	—	—	—	—	—	—	29	33	41
HARD	—	—	—	—	—	—	—	—	—	—	—	14	16	16
Robust	—	—	—	—	—	—	—	—	—	—	—	16	14	17
Terabyte	—	—	—	—	—	—	—	—	—	—	—	—	17	19
Enterprise	—	—	—	—	—	—	—	—	—	—	—	—	—	23
Spam	—	—	—	—	—	—	—	—	—	—	—	—	—	13
Participants	22	31	33	36	38	51	56	66	69	87	93	93	103	117

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 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.). Because of the technical nature of the documents, and hence the topics that could be asked against those documents, topic development and relevance judging for the enterprise track were performed by the track participants.

The track contained three search tasks: a known-item search for a particular message in the email lists archive; an ad hoc search for the set of messages that pertain to a particular discussion covered in the email lists; and a search-for-experts task. The motivation for the expert-finding task is being able to determine who the correct contact person for a particular matter is in a large organization. For the track task, the topics were the names of W3C working groups (e.g., "Web Services Choreography"), and the correct answers were assumed to be the members of that particular working group. Systems were to return the names of the people themselves, not documents that stated the people were members of the particular working group.

Twenty-three groups participated in the enterprise track, 14 groups in the discussion search task, 9 groups in the expert-finding task, and 17 groups in the known-item search task. While groups generally attempted to exploit the thread structure and quoted material in the email tasks, the effectiveness of the searches was

generally dominated by traditional content factors. Thus, more work is needed to understand how best to support discussion search.

3.2 The genomics track

The goal of genomics track is to provide a forum for evaluation of information retrieval 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 retrieval effectiveness. As in TREC 2004, the 2005 genomics track contained an ad hoc retrieval task and a categorization task.

The document set for the ad hoc task was the same corpus as was used in the 2004 genomics ad hoc task, a 10-year subset (1994 to 2003) of MEDLINE, the bibliographic database of biomedical literature maintained by the US National Library of Medicine. The corpus contains about 4.5 million MEDLINE records (which include title and abstract as well as other bibliographic information) and is about 9GB of data. The topics were developed from interviews from real biologists who were asked to fill in a "generic topic template" or GTT. The GTTs were used to produced more structured topics than traditional TREC topics so systems could make better use of resources such as ontologies and databases. The 50 test topics contain ten instances for each of the following five GTTs, where the underlined portions represent the template slots:

1. Find articles describing standard methods or protocols for doing some sort of experiment or procedure.
2. Find articles describing the role of a gene involved in a given disease.
3. Find articles describing the role of a gene in a specific biological process.
4. 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.
5. Find articles describing one or more mutations of a given gene and its biological impact.

For example, a topic derived from the mutation GTT might be *Provide information about Mutation of Ret in thyroid function.* Relevance judgments were made by assessors with backgrounds in biology using a three-point scale of definitely relevant, probably relevant, and not relevant. Both definitely relevant and probably relevant were considered relevant when computing evaluation scores.

The genomics domain has a number of model organism database projects in which the literature regarding a specific organism (such as a mouse) is tracked and annotated with the function of genes and proteins. The classification task used in the 2005 track focused on one of the tasks in this curation process, the "document triage" task. The document triage task is essentially a filtering task in which a document passes through the filter only if it should receive more careful examination with respect to a specific category. Four different categories were used in the track: Gene Ontology (GO) annotation, tumor biology, embryologic gene expression, and alleles of mutant phenotypes. The document set was the same document set used in the TREC 2004 genomics categorization task, the full text articles from a two-year span of three journals made available to the track through Highwire Press. The truth data for the task came from the actual annotation process carried out by the human annotators in the mouse genome informatics (MGI) system.

The genomics track had 41 participants, with 32 groups participating in the ad hoc search task and 19 participating in the categorization task. Retrieval effectiveness was roughly equivalent across the different topic types in the ad hoc search task. In contrast, system effectiveness was strongly dependent on the specific category in the triage task.

3.3 The HARD track

The goal of the "High Accuracy Retrieval from Documents" (HARD) track is improving retrieval system effectiveness by personalizing the search to the particular user. For the 2005 track, the method for obtaining information about the user was through clarification forms, a limited type of interaction between the system and the searcher.

The underlying task in the HARD track is an ad hoc retrieval task. Participants first submit baseline runs using the topic statements as is. They may then collect information from the searcher (the assessor who judged the topic) using clarification forms. A clarification form is a single, self-contained HTML form created by the participating group and specific to a single topic. There were no restrictions on what type of data could be collected using a clarification form, but the searcher spent no more than three minutes filling out any one form. An example use of a clarification form is to ask the searcher which of a given set of terms are likely to be good search terms for the topic. Finally, participants make new runs using the information gathered from clarification forms.

The same document set, topics, and hence relevance judgments were used in both the HARD and robust tracks. The document set was the *AQUAINT Corpus of English News Text* (LDC catalog number LDC2002T31, see `www.ldc.upenn.edu`). The 50 test topics were a subset of the topics used in previous TREC robust tracks, which had been demonstrated to be difficult topics for systems when used on the TREC disks 4&5 document set. Relevance judgments were performed by NIST assessors based on pools of both HARD and robust runs.

The motivation for sharing the test collection between the two tracks was partly financial—NIST did not have the resources to create a separate collection for each track—but sharing also had technical benefits as well. One hypothesis as to why previous years' HARD tracks did not demonstrate as large a difference in effectiveness between baseline and final runs as expected was that many of the topics in those test sets did not really need clarification. Using topics that had been shown to be difficult in the past was one way of constructing a test set that had room for improvement. The design also allows direct comparison between the largely automatic methods used in the robust track with the limited searcher feedback of the HARD track.

Sixteen groups participated in the HARD track. The majority of runs that used clarification forms did improve over their corresponding baseline runs, and a few such runs showed noticeable improvement. While this supports the hypothesis that some forms of limited user interaction can be effective in improving retrieval effectiveness, many questions regarding how best to use it remain. Note, for example, that the best automatic run from the robust track (that used no interaction) was more effective than any of the automatic runs from the HARD track.

3.4 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 main task in the TREC 2005 track was very similar to the TREC 2004 task, though there were additional tasks as well in TREC 2005.

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. 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.

The document set used in the track was again the AQUAINT corpus. The test set consisted of 75 series of questions where the target was either a person, an organization, an entity to be defined (e.g., "kudzu"), or an event. Events were new to the TREC 2005 task.

One of the concerns expressed at both the SIGIR 2004 IR4QA workshop and the QA track workshop at the TREC 2004 meeting was a desire to build infrastructure that would allow a closer examination of the role document retrieval techniques play in supporting QA technology. To this end, participants in the main task were required to submit a document ranking of the documents their system used in answering the question for each of 50 individual questions (not series). While not all QA systems produce a ranked list of documents as an initial step, some ranking (even if it consisted of only a single document) was still required. The submitted document rankings were pooled as in a traditional ad hoc task, and NIST assessors judged the pools using "contains an answer to the question" as the definition of relevant. The judged pools thus give the number of instances of correct answers in the collection, a statistic not computed for other QA test sets. The ranked lists will also support research on whether some document retrieval techniques are better than others in support of QA.

The relationship task was an optional second task in the track. The task was based on a pilot evaluation that was run in the context of the ARDA AQUAINT program (see `http://trec.nist.gov/data/qa/add_qaresources.html`). AQUAINT defined a relationship as the ability of one entity to influence another, including both the means to influence and the motivation for doing so. Eight spheres of influence were noted, including financial, movement of goods, family ties, communication pathways, organizational ties, co-location, common interests, and temporal. Systems were given a topic statement that set the context for a final question asking about one of the types of influence. The system response was a set of "information nuggets" that provided the evidence (or lack thereof) for the relationship hypothesized in the question. The relationship task test set contained 25 topics. Submissions to the relationship task were allowed to be either automatic (no manual processing at all) or manual.

Thirty-three groups participated in the main task, including three groups that performed only the document ranking task. Six groups participated in the relationship task as well. The document ranking task results demonstrated only a weak correlation between the effectiveness of the initial document ranking as measured by R-precision and the ability of the system to answer factoid questions.

3.5 The robust track

The robust track looks to improve the consistency of retrieval technology by focusing on poorly performing topics. Previous editions of the track have demonstrated that average effectiveness masks individual topic effectiveness, and that optimizing standard average effectiveness usually harms the already ineffective topics.

The task in the track is an ad hoc retrieval task where effectiveness is measured as a function of worst-case behavior. Measures of poor performance used in earlier tracks were problematic because they are relatively unstable when used with as few as 50 to 100 topics. A new measure developed during the final analysis of the TREC 2004 robust track results appears to give appropriate emphasis to poorly performing topics in addition to being stable with as few as 50 topics. This "gmap" measure is based on a geometric, rather than arithmetic, mean of average precision over a set of topics, and was the main effectiveness measure used in this year's track.

As discussed in the HARD track section, the HARD and robust tracks used the same test collection in 2005. The collection consists of the AQUAINT document set and 50 topics that had been used in previous

years' robust tracks. The 50 topics were topics that had low median effectiveness (across TREC submissions) when run against TREC disks 4&5 and are therefore considered difficult topics. The topics were selected from a larger set by choosing only those topics that had at least three relevant documents in the AQUAINT collection as judged by NIST assessors. Different assessors judged the topics this year against the AQUAINT document set from those that initially judged the topics against the disks 4&5 collection.

As in the robust 2004 track, a second requirement in the track was for systems to submit a ranked list of the topics ordered by perceived difficulty. A system assigned each topic a number from 1 to 50 where the topic assigned 1 was the topic the system believed it did best on, the topic assigned 2 was the topic the system believed it did next best on, etc. This task is motivated by the hope that systems will eventually be able to use such predictions to do topic-specific processing. The quality of a prediction is measured using the area between two curves each of which plots the MAP score computed over all topics except the run's worst X topic. X ranges from 0 (so, all topics are included) to 25 (so, the average is computed over the best half of the topics). In one curve, the worst topics are defined from the run's predictions, while in the second curve the worst topics are defined using the actual average precision scores.

Seventeen groups participated in the robust track. As in previous robust tracks, the most effective strategy was to expand queries using terms derived from resources external to the target corpus. The relative difficulty of different topics, as measured by the average score across runs, differed between the disks 4&5 collection and the AQUAINT collection.

3.6 The spam track

The spam track is a second new track in 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.

There are a number of reasons why obtaining appropriate evaluation corpora is difficult. Obviously making real email streams public is not an option because of privacy concerns. Yet creating artificial corpora is also difficult. Most of the modifications to real email streams that would protect the privacy of the recipients and senders also compromises the information used by classifiers to distinguish between ham and spam. The track addressed this problem by having several corpora, some public and some private. The track also made use of 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.

Track participants submitted their classifiers to NIST. Track coordinator Gord Cormack and his colleagues at the University of Waterloo used the jig to evaluate the submitted classifiers on the private corpora. In addition, the participants used the jig themselves to evaluate the same classifiers on the public corpora and submitted the raw results from the jig on that data back to NIST.

Several measures of the quality of a classification are reported for each combination of corpus and classifier. These measures include

ham misclassification rate: the fraction of ham messages that are misclassified as spam;

spam misclassification rate: the fraction of spam messages that are misclassified as ham;

ham/spam learning curve : error rates as a function of the number of messages processed;

ROC curve: ROC (Receiver Operating Characteristic) curve that shows the tradeoff between ham/spam misclassification rates;

ROC ham/spam tradeoff score: the area above an ROC curve. This is equivalent to the probability that the spamminess score of a random ham message equals or exceeds the spamminess score of a random spam message.

Thirteen groups participated in the spam track. In addition, the organizers ran several existing spam classifiers on the various corpora and report those results as well in the spam track section of Appendix A. On the whole, the filters were effective, though each had a misclassification rate that was observable on even the smallest corpus (8000 messages). Steady-state misclassification rates were reached quickly and were not dominated by early errors, suggesting that the filters would continue to be effective in actual use.

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 other TREC collections.

The document collection used in the track was the same collection as was used in the TREC 2004 track: 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. While smaller than a full terabyte, this collection is at least an order of magnitude greater than the next-largest TREC collection. 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 encouraged for the ad hoc task since manual runs frequently contribute unique relevant documents to the pools. The efficiency and named page tasks required completely automatic processing only.

The ad hoc retrieval task used 50 information-seeking topics created for the task by NIST assessors. While systems returned the top 10 000 documents per topic so various evaluation strategies can be investigated, pools were created from the top 100 documents per topic.

The efficiency task was an extension of the ad hoc task and 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 50 000 queries mined from web search engine logs plus the title fields of the 50 topics used in the ad hoc task. Systems returned a ranked list of the top 20 documents for each query plus reported timing statistics for processing the entire query set. To measure the effectiveness of the efficiency runs, the results for the 50 queries that corresponded to the ad hoc topic set were added to the ad hoc pools and judged by the NIST assessors during the ad hoc judging.

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.

Nineteen groups participated in the track, including 18 groups participating in the ad hoc task, 13 groups in the efficiency task, and 13 groups in the named page task. While there was a wide spread in both efficiency

and effectiveness across groups, runs submitted by the same group do demonstrate that devoting more query-processing time can increase retrieval effectiveness.

4 The Future

A significant fraction of the time of one TREC workshop is spent in planning the next TREC. Two of the TREC 2005 tracks, the HARD and robust tracks, will be discontinued as tracks in TREC 2006. A variant of the HARD track's clarification form task will continue as a subtask of the question answering track; the evaluation methodology developed in the robust track will be incorporated in other tracks with ad hoc tasks. The discontinued tracks make room for two new tracks to begin in TREC 2006. The blog track will explore information seeking behavior in the blogosphere. The goal in the legal track is to develop search technology that meets the needs of lawyers to engage in effective discovery in digital document collections.

Acknowledgements

Special thanks to the track coordinators who make the variety of different tasks addressed in TREC possible. The analysis of the pools from the HARD/robust tracks and the terabyte track was done in collaboration with Chris Buckley and Ian Soboroff.

References

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

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

[3] 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.

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

[5] 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.

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

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

[8] Ellen M. Voorhees. Overview of the TREC 2005 robust retrieval track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, 2005.

[9] 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`.

[10] 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 2005

Academia Sinica	Arizona State University (2 groups)
University of Alaska Fairbanks	Beijing University of Posts and Telecommunications
Breyer, Laird	Chinese Academy of Sciences (3 groups)
CL Research	Carnegie Mellon University (2 groups)
Coveo	CSIRO ICT Centre
California State University San Marcos	The Chinese University of Hong Kong
CRM114	Dalhousie University
DaLian University of Technology	OOO Datapark
DFKI GmbH (Saarland University)	Drexel University
Dublin City University	Ecole des Mines de Saint-Etienne
Erasmus MC	Fudan University (2 groups)
Harbin Institute of Technology	The Hong Kong Polytechnic University
Hummingbird	IBM Research Lab Haifa
IBM India Research Laboratory	IBM Almaden Research Center
IBM T.J. Watson (3 groups)	Institute for Infocomm Research
Illinois Institute of Technology	Indiana University
Institut de Recherche en Informatique de Toulouse	The Johns Hopkins University
Jozef Stefan Institute	LangPower Computing, Inc.
Language Computer Corporation	LexiClone
LowLands Team	Macquarie University
Massey University	Max-Planck Institute for Computer Science
Meiji University	Microsoft Research
Microsoft Research Asia	Microsoft Research Ltd
Massachusetts Institute of Technology	The MITRE Corporation
Monash University	National Library of Medicine - University of Maryland
National Library of Medicine (Wilbur)	National Security Agency
National Taiwan University	National University of Singapore
Oregon Health & Science University	Peking University
Pontifi cia Universidade Catolica Do Rio Grande Do Sul	Queen Mary University of London
Queens College, CUNY	Queensland University of Technology
Queen's University	RMIT University:
Rutgers University (2 groups)	Sabir Research, Inc.
SAIC OIS	Simon Fraser University
SUNY Buffalo	SUNY Stony Brook
TNO and Erasmus MC	Tokyo Institute of Technology
Tsinghua University	University of Albany
University of Amsterdam (2 groups)	University of Central Florida
University College Dublin	University of Colorado School of Medicine
University of Duisburg-Essen	U. of Edinburgh and U. of Sydney
University of Geneva	University of Glasgow
University of Illinois at Chicago	University of Illinois at Urbana-Champaign
University of Iowa	University of Limerick
University of Magdeburg	University of Maryland
University of Massachusetts	The University of Melbourne
The University of Michigan-Dearborn	Universit degli Studi di Milano
University of North Carolina	Universite de Neuchatel
University of North Texas	University of Padova
Universite Paris-Sud (2 groups)	Universitat Politcnica de Catalunya
University of Pisa	University of Pittsburgh
University of Sheffi eld	University of Strathclyde
University of Tampere	The University of Tokyo
University of Twente	University of Waterloo (2 groups)
University of Wisconsin	York University

Overview of the TREC-2005 Enterprise Track

Nick Craswell
MSR Cambridge, UK
nickcr@microsoft.com

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

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

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 organisations, 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 this year 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 organisation.

Obviously, it's hard to imagine that any organisation would be willing to open its intranet to public distribution, even for research, so for the initial document collection we looked to an organisation that conducts most if not all of its day-to-day business on the public web: the World Wide Web Consortium (W3C). The collection is a crawl of the public W3C (*.w3.org) sites in June 2004. It is not a comprehensive crawl, but rather represents a significant proportion of the public W3C documents. It comprises 331,037 documents, retrieved via multithreaded breadth-first crawling. Some details of the corpus are in Table 1.

The majority of the documents in this collection are email, and thus the tasks this year focus on email. Note that the documents are not in native formats, but are rendered into HTML.

There are two tasks with a total of three experiments:

- Email search task: Using pages from `lists.w3.org`.
  - Known item experiment: 125 queries. The user is searching for a particular message, enters a query and will be satisfied if the message is retrieved at or near rank one. There were an additional 25 queries for use in training.
  - Discussion search experiment: 59 queries. The user is searching to see how pros and cons of an argument/discussion were recorded in email. Their query describes the topic, and they care both whether the results are relevant and whether they contain a pro/con. There were no training queries, and indeed no judgements prior to submission.

Table 1: Details of the W3C corpus. Scope is the name of the subcollection and also the hostname where the pages were found, for example lists.w3.org. The exception is the subcollection 'other' which contains several small hosts.

Type	Scope	Size (GB)	Docs	avdocsize (KB)
Email	lists	1.855	198,394	9.8
Code	dev	2.578	62,509	43.2
Web	www	1.043	45,975	23.8
Wiki web	esw	0.181	19,605	9.7
Misc	*other*	0.047	3,538	14.1
Web	people	0.003	1,016	3.6
	all	5.7	331,037	18.1

- Expert search task: 50 queries. Given a topical query, find a list of W3C people who are experts in that topic area. Finding people, not documents, based on analysis of the entire W3C corpus. Participants were provided with a list of 1092 candidate experts for use on all queries. There were 10 training queries.

2 Email search task

This task focuses on searching the 198,394 pages crawled from lists.w3.org. These are html-ised archives of mailing lists, 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. Some participants made their extracted information available to the group.

In the known item search experiment, participants developed (query, docno) pairs that represent a user who enters a query in order to find a specific message (item). Of the 150 pairs developed, 25 were provided for training and 125 were used for the evaluation reported here. Results are in Table 2. The measures for this task were the mean reciprocal rank (MRR) of the correct answer, and the fraction of topics with the correct answer somewhere in the top 10 ("Success at 10" or S@10). Also reported is the fraction of topics that found the correct answer anywhere in the ranking (S@inf). In recent Web Track homepage finding experiments, it was possible to find the correct homepage with $MRR > 0.7$ and $S@10 \simeq 0.9$. Known item email search results are quite good for a first year, being about 0.1 lower on both metrics.

Nearly every group took a different approach at integrating the email text with email metadata and the larger thread structure. To give some examples, University of Glasgow (uog) combined priors for web-specific features — anchor text, titles of pages — with email-specific priors — threads and dates in messages and topics [7]. Microsoft Cambridge (MSRC) used their fielded BM25 with message fields, text, and thread features [4]. CMU (CMU) mixed language models for individual messages, message subjects, threads, and subthreads, and used thread-depth priors [8]. While the initial results are encouraging, it's clear that with this many types of data to balance, more work remains to be done.

Run	MRR	S@10	S@inf
uogEDates2	**0.621**	0.784	0.920
MSRCKI5	0.613	0.816	**0.952**
covKIRun3	0.605	0.792	0.896
humEK05t3l	0.604	0.808	0.912
CMUnoPS	0.601	0.816	0.912
CMUnoprior	0.598	**0.824**	0.912
qdWcEst	0.579	0.792	0.920
priski4	0.551	0.728	0.896
KITRANS	0.536	0.728	0.880
WIMent01	0.533	0.784	0.912
csiroanuki5	0.522	0.776	0.888
UWATEntKI	0.519	0.712	0.888
csusm2	0.510	0.712	0.792
qmirkidtu	0.367	0.600	0.768
LPC5	0.343	0.480	0.504
PITTKIA1W8	0.335	0.496	0.808
LMplaintext	0.326	0.544	0.704
DrexelKI05b	0.195	0.376	0.624

Table 2: Known item results, the run from each of the 17 groups with the best MRR, sorted by MRR. The best in each column is highlighted. (An extra line was added to show the run with best S@10.)

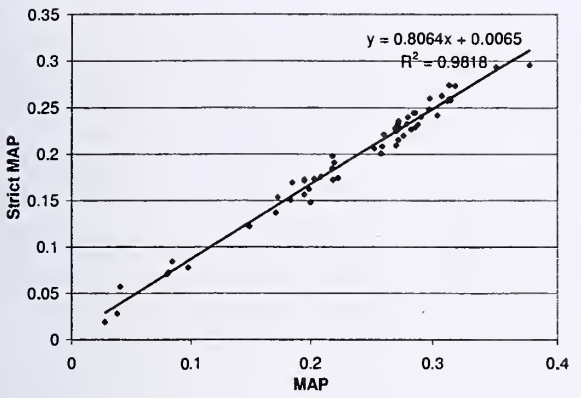

Figure 1: MAP for the 57 discussion search runs, calculated by conflating the top two (MAP) or bottom two (Strict MAP) judging levels.

In the discussion search experiment, participants developed topic descriptions and performed relevance judgements as described in Section 4. There are three types of answers: irrelevant, relevant without pro/con statement (also called "partially relevant") and relevant with pro/con statement. Table 3 shows discussion search results where any document that is not judged irrelevant is relevant (conflating the two positive judging levels). Interestingly, the top two runs are significantly better than the rest on our main measure mean average precision (MAP). For TITLETRANS, this is primarily due to the influence of a single topic [6]. The table also reports several other measures: R-precision (precision at rank R, where R is the number of relevant documents for that topic), bpref [2], precision at ranks (5, 10, 20, 30, 100, 1000), and reciprocal rank of the first relevant document retrieved.

Table 4 shows similar results if we now conflate the lower two judging levels, giving a 'strict' evaluation that only counts documents that include a pro/con statement as relevant. The overall rankings of systems are nearly identical, with a Kendall's tau of 0.893. Figure 1, shows a scatter plot, with the two types of MAP being strongly correlated.

The common focus of most groups in the discussion search subtask was how to effectively exploit thread structure and quoted material. University of Maryland (TITLETRANS in table 3 and 4) explored expanding documents using threads and the trade-off between reinforcing quoted passages and removing them altogether, with mixed results [6]. University of Amsterdam (ToNsBs) applied a straightforward language model with a filter to eliminate non-email documents [1]. Microsoft Research's (MSRC) best-performing run used only textual fields from the messages and no static features (year of message, number of parents in the thread) [4]. So it seems that these results represent mostly topic-relevance retrieval effectiveness, and we have not yet found definitive solutions to discussion search.

An important point raised by the University of Maryland team is that some of the topics did not necessarily lend themselves to pro/con arguments on the subject. Additionally, while the relevance judgements do indicate whether a pro/con argument is present in the message, we did not collect whether the argument was for or against the subject. They also found that some topics were not only more amenable to pro/con discussions, but also exhibited greater agreement between assessors. For the 2006 track, we plan to focus more closely on the topic creation process.

3 Expert search task

In the expert search task, participants could use all 331,037 documents in order to rank a list of 1092 candidate experts. This could involve creating a document for each candidate and applying simple IR techniques, or could involve natural language processing and information extraction technologies targeted at different document types such as email. Results are presented in Table 5.

For this year's pilot of this task, the search topics were so-called "working groups" of the W3C, and the experts were members of these groups. These ground-truth lists were not part of the collection but were located after the crawl was performed. This enabled us to dry-run this task with minimal effort in creating relevance judgments.

Top-scoring runs used quite advanced techniques:

THUENT0505 This run makes use of all w3c web part information and Email lists (the list part) together with inlink anchor text of these files. Text content are

Run	MAP	r-prec	bpref	P@5	P@10	p@20	p@30	P@100	P@1000	RR1
TITLETRANS	**0.3782**	**0.4051**	**0.3781**	0.5831	0.5000	0.4246	0.3712	**0.2427**	**0.0469**	0.7637
ToNsBs350F	0.3518	0.3769	0.3588	0.5729	**0.5407**	**0.4449**	**0.3768**	0.2147	0.0439	**0.7880**
UwatEntDSq	0.3187	0.3514	0.3266	0.5153	0.4831	0.4034	0.3610	0.2244	0.0415	0.6860
csiroanuds1	0.3148	0.3597	0.3310	0.5593	0.5102	0.4051	0.3469	0.2037	0.0416	0.7292
MSRCDS2	0.3139	0.3583	0.3315	**0.5864**	0.5169	0.4127	0.3475	0.1966	0.0428	0.7423
irmdLTF	0.3138	0.3461	0.3318	0.5254	0.4797	0.4169	0.3729	0.2183	0.0409	0.7249
prisds1	0.3077	0.3393	0.3294	0.5797	0.4966	0.3881	0.3277	0.1815	0.0381	0.6617
du05quotstrg	0.2978	0.3431	0.3163	0.5288	0.4712	0.3881	0.3362	0.2047	0.0417	0.6793
qmirdju	0.2860	0.3202	0.3017	0.5119	0.4695	0.3788	0.3226	0.1976	0.0421	0.7026
LMlam08Thr	0.2721	0.3062	0.2884	0.3932	0.3746	0.3263	0.2887	0.1819	0.0412	0.5678
PITTDTA2SML1	0.2184	0.2494	0.2333	0.3864	0.3271	0.2712	0.2288	0.1339	0.0290	0.4759
MU05ENd5	0.2182	0.2655	0.2530	0.4407	0.3831	0.3136	0.2893	0.1819	0.0381	0.6121
NON	0.0843	0.1305	0.1082	0.2576	0.2237	0.1771	0.1508	0.0869	0.0087	0.4123
LPC1	0.0808	0.0981	0.0907	0.2237	0.1746	0.1305	0.1062	0.0544	0.0072	0.3670

Table 3: Discussion search: Evaluation where judging levels 1 and 2 are 'relevant'. Lists the run with best MAP from each of the 14 groups, sorted by MAP. The best in each column is highlighted.

Run	MAP	r-prec	bpref	P@5	P@10	p@20	p@30	P@100	P@1000	RR1
TITLETRANS	**0.2958**	0.3064	**0.3381**	0.3661	0.3356	0.2797	0.2429	**0.1531**	**0.0279**	0.5710
ToNsBs350F	0.2936	**0.3065**	0.3286	0.4068	**0.3763**	**0.2907**	0.2407	0.1292	0.0256	0.6247
MSRCDS2	0.2742	0.2892	0.3043	**0.4339**	0.3661	0.2864	0.2282	0.1200	0.0253	**0.6376**
UwatEntDSq	0.2735	0.2990	0.3086	0.3593	0.3220	0.2669	0.2373	0.1388	0.0250	0.5612
prisds1	0.2626	0.2803	0.2977	0.4000	0.3407	0.2695	0.2232	0.1136	0.0237	0.5234
du05quotstrg	0.2600	0.2837	0.2883	0.3864	0.3356	0.2576	0.2226	0.1246	0.0246	0.5436
irmdLTF	0.2592	0.2712	0.2852	0.3966	0.3407	0.2881	**0.2514**	0.1464	0.0247	0.5890
csiroanuds1	0.2583	0.2854	0.3000	0.3864	0.3492	0.2712	0.2243	0.1253	0.0253	0.5791
qmirdju	0.2446	0.2750	0.2841	0.3492	0.3153	0.2568	0.2085	0.1236	0.0248	0.5673
LMlam08Thr	0.2153	0.2442	0.2409	0.2576	0.2390	0.2068	0.1836	0.1149	0.0254	0.4369
PITTDTA2SML1	0.1978	0.2072	0.2165	0.2949	0.2508	0.1907	0.1565	0.0868	0.0176	0.4110
MU05ENd5	0.1847	0.2262	0.2309	0.3322	0.2627	0.2136	0.1989	0.1214	0.0230	0.5518
NON	0.0842	0.1285	0.1099	0.1864	0.1678	0.1280	0.1040	0.0568	0.0057	0.3061
LPC1	0.0724	0.0872	0.0811	0.1661	0.1220	0.0873	0.0723	0.0369	0.0050	0.3012

Table 4: Discussion search: Strict evaluation, where only judging level (includes a pro/con statement) is considered relevant. Lists the run with best MAP from each of the 14 groups, sorted by MAP. The best in each column is highlighted.

reconstructed and formed description files for each candidate person. Structure information inside web pages was also used to improve performance. Words from important pages are emphasised in this run. Bigram retrieval was also applied [5].

MSRA054 The basic model plus cluster-based reranking. (The basic model, 1) a two-stage model of combining relevance and co-occurrence 2) the co-occurrence model consists of body-body, title-author, and title-tree submodels 3) a back-off query term matching method which prefers exact match, then partial match, and finally word-level match.) [3]

This suggests that there were gains in effectiveness to be had via leveraging the heterogeneity of the dataset and the 'information extraction' flavor of the task. On the other hand, some groups (including THU and others) did notice that the search topics were W3C working groups, and took advantage of this fact by mining working group membership knowledge out of the collection. Thus, these results should be considered preliminary pending a more realistic expert search data set.

4 Judging

Since each known item topic is developed with a particular message in mind, that message is by definition the only answer needed, so no further relevance judging is required. However, in a corpus with significant duplication, it may be necessary to examine the pool for duplicates or near-duplicates of the item, as in the Web and Terabyte tracks. This year, because we do not believe that duplication is such a problem in `lists.w3.org`, we decided to expend effort in duplicate identification, so each query has exactly one answer.

Similarly, there was no judging required for the expert search task. This is because we used working group membership as our ground truth, as described in Section 3.

For the discussion search task, the judging was more involved. Because it is an ad hoc search task, it needs true relevance judgments, but the technical nature of the collection meant that NIST assessors would not be ideal topic creators or relevance judges. Instead, track participants both created the topics and judged the pools to determine the final relevance judgments.

In response to a call for participation in April, thirteen groups submitted candidate topics for the discussion search and known item tasks. For the known item search task, the topics included the query/name for the page and the target docno. For discussion search, the topic included a "query" field (equivalent to the traditional "title" field) and a "narrative" field to delineate the relevance boundary of the topic. In all, 63 topics were submitted, and NIST selected 60 topics for the final set.

Judging was done over the internet using an assessment system at CWI. Each topic was assigned to two groups, the group who authored the topic (the primary assessor) and another group (the secondary assessor). Secondary assessment assignments were made so as to balance authors across judging groups and to somewhat limit overall judging load. The topics and judging groups are shown in table 6. One group created three topics (24, 27, and 46) but did not submit any runs or respond to requests to help judge; their topics were reassigned to groups A, B, and C respectively as primary judges. Groups M and N did not contribute topics but did submit runs and agreed to help judge as secondary assessors. The pools were intentionally kept small to reduce the judging burden on sites. Three runs from each group were pooled to a depth of 50, and the final pools contained between 249 and 865 documents (mean 529).

Judging began in August and ran through early October, and was extremely successful, with all but three topics fully judged by their primary assessor, and 52 by the secondary assessor. The official qrels set consists of the primary judgments for 56 topics, and the secondary judgments for the remaining topics (26, 53, and 57). No relevant documents were found by the primary assessor for topic 4, and so we have left this topic out. This qrels set contains 31,258 judgments: 27,813 irrelevant, 1,441 relevant non-pro/con (R1) and 2,004 relevant pro/con (R2) messages. Median per topic was 14 for R1 and 20 for R2.

At the time of this writing, we have done some examination of the affects of assessor disagreement, by comparing the ranking of systems according to the primary and secondary judgments. For this experiment, we considered the 48 topics for which judgments exist from both assessors (and again dropping topic 4). Comparing the rankings of systems using each set of judgments yields a Kendall's tau of 0.763, which is less than the level of 0.9 taken to indicate "essentially identical", but still signifi-

Run	MAP	r-prec	bpref	P@5	P@10	P@20	P@30	P@100	P@1000	RR1
THUENT0505	**0.2749**	**0.3330**	0.4880	**0.4880**	**0.4520**	**0.3390**	**0.2800**	0.1142	0.0114	**0.7268**
MSRA054	0.2688	0.3192	**0.5685**	0.4080	0.3700	0.3190	0.2753	0.1306	**0.0131**	0.6244
MSRA055	0.2600	0.3089	0.5655	0.3920	0.3580	0.3150	0.2733	**0.1308**	**0.0131**	0.5832
CNDS04LC	0.2174	0.2631	0.4299	0.4120	0.3460	0.2820	0.2240	0.0942	0.0094	0.6068
uogES05CbiH	0.1851	0.2397	0.4662	0.3800	0.3160	0.2600	0.2133	0.1130	0.0113	0.5519
PRISEX3	0.1833	0.2269	0.4182	0.3440	0.3080	0.2530	0.2087	0.1026	0.0103	0.5614
uams05run1	0.1277	0.1811	0.3925	0.2720	0.2220	0.2000	0.1753	0.0944	0.0094	0.4380
DREXEXP1	0.1262	0.1743	0.3409	0.3120	0.2500	0.1760	0.1467	0.0720	0.0072	0.4635
LLEXemails	0.0960	0.1357	0.2985	0.2000	0.1860	0.1530	0.1213	0.0628	0.0063	0.4054
qmirex4	0.0959	0.1511	0.2730	0.2360	0.1880	0.1390	0.1233	0.0534	0.0053	0.4189

Table 5: Expert search results, the run from each of the 9 groups with the best MAP, sorted by MAP. The best in each column is highlighted. (An extra line was added to show the run with best P@100.)

Group	Authored topics	Assigned topics	Total
A	7 8 33 41 52 24	12 25 48 60	10
B	4 37 43 51 60 27	13 26 49	9
C	6 11 20 34 48 46	14 37 50	9
D	9 19 58	1 15 27 38 51	8
E	3 15 23 31 35	2 16 28 39 52	10
F	5 10 14 16 36	3 17 29 40 53	10
G	1 2 25 26 53	4 18 41 54	9
H	39 40 50 56	5 19 30 42 55	9
I	18 30 45	6 31 36 43 56	8
J	12 32 47 55 57	7 20 44 46	9
K	22 29 38 42 49	8 21 32 45	9
L	13 17 21 28 44 54 59	9 22 33 57	11
M		10 23 34 47 58	5
N		11 24 35 59	4

Table 6: Topic assignments for relevance assessment. "Authored topics" were created by that group. "Assigned topics" were assigned to that group by NIST for judging.

cantly correlated ($p < 2.2 \cdot 10^{16}$). We intend to look more closely at this data to see if particular topics or assessors cause more variation in the ranking.

5 Conclusion

This year participants made heavy use of email structure and combination of evidence techniques in email search and expert search with some success, but there remains much to learn. In future enterprise search experiments it would be nice to further our exploration of novel data types such as email archives, and of novel tasks such as expert search. This might include incorporation of a greater amount of real user data (perhaps query and click logs) to enhance our focus on enterprise user tasks.

For discussion search, we plan to approach topic creation with more care. Specifically, next year's topics will more closely target pro/con discussions, and we may ask assessors to label messages as either pro, con, both, or can't tell.

This year's foray into community-developed topics and relevance judgments marked a significant change for TREC, although such is the practise in other forums such as INEX. It has been a very successful experience, and we intend to continue collection development this way next year.

Task details for this year are maintained on the track wiki, at `http://www.ins.cwi.nl/projects/trec-ent/wiki/index.php/Main_Page`.

Acknowledgements

We are grateful to the World Wide Web Consortium for allowing us to make a snapshot of their site available as a research tool. We also thank the University of Glasgow for hosting a Terrier-based search interface to the W3C collection for topic development. Lastly, we thank the participants of the 2005 enterprise track for helping to create the test collection.

References

[1] Leif Azzopardi, Krisztian Balog, and Maarten de Rijke. Language modeling approaches for enterprise tasks. In Voorhees and Buckland [9].

[2] 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 (SIGIR 2004)*, pages 25–32, Sheffield, UK, July 2004. ACM Press.

[3] Yunbo Cao, Jingjing Liu, Shenghua Bao, and Hang Li. Research on expert search at enterprise track of TREC 2005. In Voorhees and Buckland [9].

[4] Nick Craswell, Hugo Zaragoza, and Stephen Robertson. Microsoft Cambridge at TREC-14: Enterprise track. In Voorhees and Buckland [9].

[5] Yupeng Fu, Wei Yu, Yize Li, Yiqun Liu, Min Zhang, and Shaoping Ma. THUIR at TREC 2005: Enterprise track. In Voorhees and Buckland [9].

[6] Jimmy Lin, Eileen Abels, Dina Demner-Fushman, Douglas W. Oard, Philip Wu, and Yejun Wu. A menagerie of tracks at maryland: HARD, enteprise, QA, and genomics, oh my! In Voorhees and Buckland [9].

[7] Craig Macdonald, Ben He, Vassilis Plachouras, and Iadh Ounis. University of Glasgow at TREC 2005: Experiments in terabyte and enterprise tracks with terrier. In Voorhees and Buckland [9].

[8] Paul Ogilvie and Jamie Callan. Experiments with language models for known-item finding of e-mail messages. In Voorhees and Buckland [9].

[9] E. M. Voorhees and L. P. Buckland, editors. *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, November 2004.

TREC 2005 Genomics Track Overview

William Hersh[1], Aaron Cohen[1], Jianji Yang[1], Ravi Teja Bhupatiraju[1],
Phoebe Roberts[2], Marti Hearst[3]
[1]Oregon Health & Science University, Portland, OR, USA
[2]Biogen Idec Corp., Boston, MA, USA
[3]University of California, Berkeley, CA, USA

The TREC 2005 Genomics Track featured two tasks, an ad hoc retrieval task and four subtasks in text categorization. The ad hoc retrieval task utilized a 10-year, 4.5-million document subset of the MEDLINE bibliographic database, with 50 topics conforming to five generic topic types. The categorization task used a full-text document collection with training and test sets consisting of about 6,000 biomedical journal articles each. Participants aimed to triage the documents into categories representing data resources in the Mouse Genome Informatics database, with performance assessed via a utility measure.

1. Introduction

The goal of the TREC Genomics Track is to create test collections for evaluation of information retrieval (IR) and related tasks in the genomics domain. The Genomics Track differs from other TREC tracks in that it is focused on retrieval in a specific domain as opposed to general retrieval tasks, such as Web searching or question answering. There are many reasons why a focus on this domain is important. New advances in biotechnologies have changed the face of biological research, particularly "high-throughput" techniques such as gene microarrays [1]. These techniques not only generate massive amounts of data but also have led to an explosion of new scientific knowledge. As a result, this domain is ripe for improved information access and management.

The scientific literature plays a key role in the growth of biomedical research data and knowledge. Experiments identify new genes, diseases, and other biological processes and factors that require further investigation. Furthermore, the literature itself becomes a source of "experiments" as researchers turn to it to search for knowledge that in turn drives new hypotheses and research. Thus, there are considerable challenges not only for better IR systems, but also for improvements in related techniques, such as information extraction and text mining [2, 3].

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

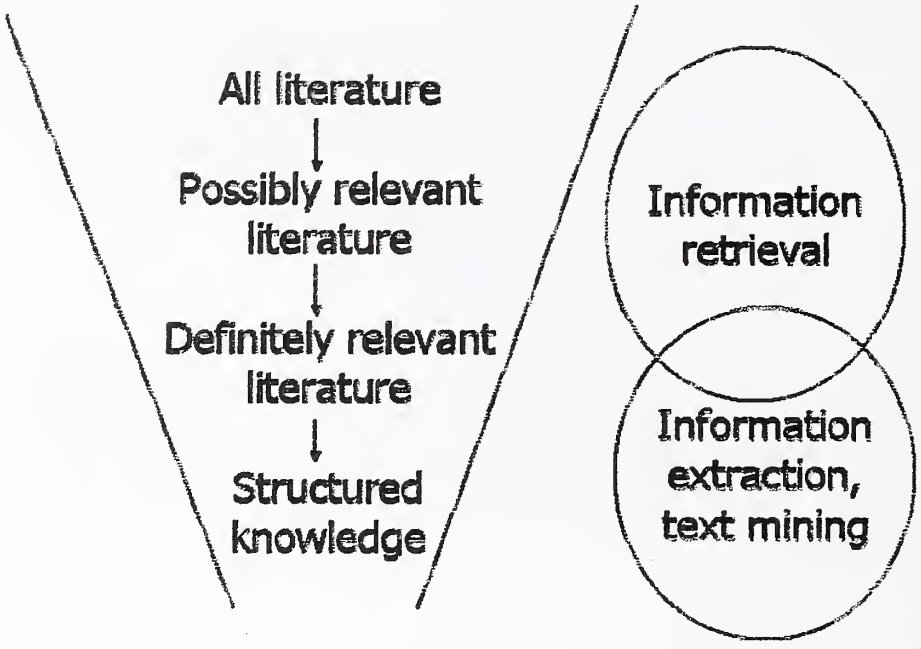

Figure 1 - The funneling of scientific literature and related information retrieval and extraction disciplines.

TREC 2005 marks the third offering of the Genomics Track. The first of the track, 2003, was limited by lack of resources to perform relevance judgments and other tasks, so the track had to use "pseudojudgments" culled from data created for other purposes [4]. In 2004, however, the track obtained a five-year grant from the U.S. National Science Foundation (NSF), which provided resources for building test collections and other data sources. The 2004 track featured an ad hoc retrieval task [5] and three subtasks in text categorization [6].

For 2005, the track built on the success of 2004 by using the same underlying document collections on new topics for ad hoc retrieval and refinement of the text categorization tasks. Similar to the 2004 track, the track attracted the largest number of participating groups of any in TREC. In 2005, 32 groups submitted 59 runs to the ad hoc retrieval task, while 19 groups submitted 192 runs to the categorization subtasks. A total of 41 different groups participated, with 10 groups participating in both tasks, 22 participating only in the ad hoc retrieval task, and 9 participating in just the categorization tasks, making it the largest track in TREC 2005.

The remainder of this paper covers the tasks, methods, and results of the two tasks separately, followed by discussion of future directions.

2. Ad Hoc Task

2.1 Task

The ad hoc retrieval task modeled the situation of a user with an information need using an information retrieval system to access the biomedical scientific literature. The document collection was based on a large subset of the MEDLINE bibliographic database. It should be

noted that although we are in an era of readily available full-text journals (usually requiring a subscription), many users of the biomedical literature enter through searching MEDLINE. As such, there are still strong motivations to improve the effectiveness of searching MEDLINE.

2.2 Documents

The document collection for the 2005 ad hoc retrieval task was the same 10-year MEDLINE subset using for the 2004 track. One goal we have is to produce a number of topic and relevance judgment collections that use this same document collection to make retrieval experimentation easier (so people do not have to load different collections into their systems). Additional uses of this subset have already appeared [7]. MEDLINE can be searched by anyone in the world using the PubMed system of the National Library of Medicine (NLM), which maintains both MEDLINE and PubMed. The full MEDLINE database contains over 14 million references dating back to 1966 and is updated on a daily basis.

The subset of MEDLINE for the TREC 2005 Genomics Track consisted of 10 years of completed citations from the database inclusive from 1994 to 2003. Records were extracted using the Date Completed (DCOM) field for all references in the range of 19940101 - 20031231. This provided a total of 4,591,008 records, which is about one third of the full MEDLINE database. The data included all of the PubMed fields identified in the MEDLINE Baseline record. Descriptions of the various fields of MEDLINE are available at: http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#MEDLINEDisplayFormat

The MEDLINE subset was provided in the "MEDLINE" format, consisting of ASCII text with fields indicated and delimited by 2-4 character abbreviations. The size of the file uncompressed was 9,587,370,116 bytes. An XML version of MEDLINE subset was also available. It should also be noted that not all MEDLINE records have abstracts, usually because the article itself does not have an abstract. In general, about 75% of MEDLINE records have abstracts. In our subset, there were 1,209,243 (26.3%) records without abstracts.

2.3 Topics

As with 2004, we collected information needs from real biologists. However, instead of soliciting free-form biomedical questions, we developed a set of six generic topic templates (GTTs) derived from an analysis of the topics from the 2004 track and other known biologist information needs (Table 1). GTTs consist of semantic types, such as genes or diseases, placed in the context of commonly queried biomedical questions, and semantic types are often present in more than one GTT. After we developed the GTTs, 11 people interviewed 25 biologists to obtain ten or more specific information needs that conformed to each GTT. One GTT did not model a commonly researched problem, and was dropped from the study. The topics did not have to fit precisely into the GTTs, but had to come close, i.e., have all the required semantic types. We then had other people search on the topics to make sure there was some, but not too much, relevant information in MEDLINE.). Ten information needs for each GTT were selected for inclusion in the 2005 track to total fifty topics.

In order to get participating groups started with the topics, and in order for them not to "spoil"

their automatic status of their official runs by working with the official topics, we developed 10 sample topics, consisting of two topics from each GTT. These learning topics had a MEDLINE search and relevance judgments of the output that we made available to participants. Table 1 also gives an example topic for each GTT that comes from the sample topics.

2.4 Relevance judgments

Relevance judgments were done using the conventional pooling method of TREC. Based on estimation of relevance judgment resources, the top 60 documents for each topic from all official runs were used. This gave an average pool size of 821 documents with a range of 290 to 1356. These pools were then provided to the relevance judges, who consisted of five individuals with varying expertise in biology. The relevance judges were instructed in the following manner for each GTT:

- Relevant article must describe how to conduct, adjust, or improve a standard, a, new method, or a protocol for doing some sort of experiment or procedure.
- Relevant article must describe some specific role of the gene in the stated disease or biological process.
- Relevant article must describe a specific interaction (e.g., promote, suppress, inhibit, etc.) between two or more genes in the stated function of the organ or the disease.
- Relevant article must describe a mutation of the stated gene and the particular biological impact(s) that the mutation has been found to have.

The articles had to describe a specific gene, disease, impact, mutation, etc. and not just the concept in general.

Table 1 - Generic topic types and example sample topics. The semantic types in each GTT are underlined.

Generic Topic Type	**Topic Range**	**Example Sample Topic**
Find articles describing standard methods or protocols for doing some sort of experiment or procedure	100-109	Method or protocol: GST fusion protein expression in Sf9 insect cells
Find articles describing the role of a gene involved in a given disease	110-119	Gene: DRD4 Disease: Alcoholism
Find articles describing the role of a gene in a specific biological process	120-129	Gene: Insulin receptor gene Biological process: Signaling 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	130-139	Genes: HMG and HMGB1 Disease: Hepatitis
Find articles describing one or more mutations of a given gene and its biological impact	140-149	Gene with mutation: Ret Biological impact: Thyroid function

Relevance judges were asked to rate documents as definitely, possibly, or not relevant. As in 2004, articles that were rated definitely or possibly relevant were considered relevant for use in the binary recall and precision-related measures of retrieval performance. Relevance judgments were performed by individuals with varying levels of expertise in biology (from an undergraduate student to a PhD researcher). For 10 of the topics, judgments were performed in duplicate to allow interobserver reliability measurement using the kappa statistic.

2.5 Measures and statistical analysis

Retrieval performance was measured with the "usual" TREC ad hoc measures of mean average precision (MAP), binary preference (B-Pref) [8], precision at the point of the number of relevant documents retrieved (R-Prec), and precision at varying numbers of documents retrieved (e.g., 5, 10, 30, etc. documents up to 1,000). These measures were calculated using version 8.0 of trec_eval developed by Chris Buckley (Sabir Research).

Research groups submitted their runs through the TREC Web site in the usual manner. They were required to classify their runs into one of three categories:

- Automatic - no manual intervention in building queries
- Manual - manual construction of queries but no further human interaction
- Interactive - completely interactive construction of queries and further interaction with system output

They were also required to provide a brief system description.

Statistical analysis of the above measures was performed using SPSS (version 12.0). Repeated measure analysis of variance (ANOVA) with posthoc tests using Sidak adjustments were performed on the above variables. In addition, descriptive analysis of MAP was also done to study the spread of the data.

2.6 Results

A total of 32 groups submitted 58 runs. Table 2 shows the results of relevance judging for each topic, listing the pool size sent to a given assessor plus their distribution of relevance assessments. The combined number and percentage of documents rated definitely and possibly relevant are also listed, since these were considered relevant from the standpoint of official results. Six topics had no definitely relevant documents. One topic had no definitely or possibly relevant documents and was dropped from the calculation of official results.

Table 2 - Relevant documents per topic. Topic 135 had no relevant documents and was eliminated from the results. Documents that were definitely or possibly relevant were considered to be relevant for the purposes of official TREC results.

Topic	Pool Size	Definitely Relevant	Possibly Relevant	Not Relevant	Definitely + Possibly (TREC) Relevant	% TREC Relevant
100	704	22	52	630	74	10.5%
101	651	2	18	631	20	3.1%
102	1164	5	5	1154	10	0.9%
103	701	6	19	676	25	3.6%
104	629	0	4	625	4	0.6%
105	1133	4	85	1044	89	7.9%
106	1230	44	125	1061	169	13.7%
107	484	76	114	294	190	39.3%
108	1092	76	127	889	203	18.6%
109	389	165	14	210	179	46.0%
110	934	4	12	918	16	1.7%
111	675	109	93	473	202	29.9%
112	872	4	7	861	11	1.3%
113	1356	10	4	1342	14	1.0%
114	754	210	169	375	379	50.3%
115	1350	3	12	1335	15	1.1%
116	1265	58	28	1179	86	6.8%
117	1094	527	182	385	709	64.8%
118	938	20	12	906	32	3.4%
119	589	42	19	528	61	10.4%
120	527	223	122	182	345	65.5%
121	422	17	25	380	42	10.0%
122	871	19	37	815	56	6.4%
123	1029	5	32	992	37	3.6%
124	752	8	53	691	61	8.1%
125	1202	3	8	1191	11	0.9%
126	1320	190	117	1013	307	23.3%
127	841	1	3	837	4	0.5%
128	954	21	53	880	74	7.8%
129	987	16	22	949	38	3.9%
130	813	9	23	781	32	3.9%
131	431	2	40	389	42	9.7%
132	531	3	27	501	30	5.6%
133	523	0	5	518	5	1.0%
134	732	2	9	721	11	1.5%
135	1057	0	0	1057	0	0.0%
136	853	0	3	850	3	0.4%
137	1129	12	39	1078	51	4.5%
138	501	6	6	489	12	2.4%
139	380	15	20	345	35	9.2%
140	395	14	15	366	29	7.3%
141	520	34	47	439	81	15.6%
142	528	151	120	257	271	51.3%
143	902	0	4	898	4	0.4%
144	1212	1	1	1210	2	0.2%
145	288	10	22	256	32	11.1%
146	825	370	67	388	437	53.0%
147	659	0	10	649	10	1.5%
148	536	0	11	525	11	2.1%
149	1294	6	17	1271	23	1.8%
Avg	820.4	50.5	41.2	728.7	91.7	12.5%

Table 3 - Overlap of duplicate judgments for kappa statistic.

	Duplicate judge - Relevant	**Duplicate judge - Not Relevant**	**Total**
Original judge - Relevant	1100	629	1729
Original judge - Not Relevant	546	8204	8750
Total	1646	8833	10479

In order to assess the consistency of relevance judgments, we had judgments of ten topics performed in duplicate. (For three topics, we actually had judgments performed in triplicate; one of these was the topic that had no relevant documents.) The judgments from the original judge who did the assessing was used as the "official" judgment. Table 3 shows the consistency of the judgments from the original and duplicating judge. The kappa score for inter-judge agreement was 0.585, indicating a "moderate" level of agreement and comparable to the 2004 Genomics Track.

The overall results are shown in Table 4, sorted by MAP. The top-ranking run came from York University. The top-ranking run was a manual run, but this group also had the top-ranking automatic run. The top-ranking interactive run was somewhat further down the list, although this group had an automatic run that performed better. The statistical analysis of the runs showed overall statistical significance for all of the measures. Pair-wise comparison of MAP for the 58 runs showed that significant difference from the top run was obtained at run uta05i. At the other end, significant difference from the lowest run was reached by run genome2. Figure 2 shows the MAP results with 95% confidence intervals, while Figure 3 shows all of the statistics from Table 4, sorted by each run's MAP.

We also assessed the results by topic. Table 5 shows the various measures for each topic, while Figure 4 shows the same data graphically with confidence intervals. The spread of MAP showed a wide variation among the 49 topics. Topic 136 had the lowest variance (<0.001) with range of 0-0.0287. On the other hand, topic 119 showed the highest variance (0.060), with range of 0.0144-0.8289. Topic 121 received the highest mean MAP at 0.620, while topic 143 had the lowest at 0.003. Figure 5 compares the number of relevant documents with MAP for each topic.

In addition, we grouped the results by GTT, as shown in Table 6. The GTT of information describing the role of a gene in a disease achieved the highest MAP, while the gene interactions and gene mutations achieved the best B-Pref. However, the differences among all of the GTTs were modest.

Table 4 - Run results by run name, type (manual, automatic, or interactive), and performance measures.

Run	Group	Type	MAP	R-Prec	B-pref	P10	P100	P1000
york05gm1 [9]	yorku.huang	m	0.302	0.3212	0.3155	0.4551	0.2543	0.0748
york05ga1 [9]	yorku.huang	a	0.2888	0.3118	0.3061	0.4592	0.2557	0.0721
ibmadz05us [10]	ibm.zhang	a	0.2883	0.3091	0.3026	0.4735	0.2643	0.0766
ibmadz05bs [10]	ibm.zhang	a	0.2859	0.3061	0.2987	0.4694	0.2606	0.0761
uwmtEg05	uwaterloo.clarke	a	0.258	0.2853	0.2781	0.4143	0.2292	0.0718
UIUCgAuto [11]	uiuc.zhai	a	0.2577	0.2688	0.2708	0.4122	0.231	0.0709
UIUCgInt [11]	uiuc.zhai	i	0.2487	0.2627	0.267	0.4224	0.2355	0.0694
NLMfusionA [12]	nlm-umd.aronson	a	0.2479	0.2767	0.2675	0.402	0.2378	0.0688
iasl1 [13]	academia.sinica.tsai	a	0.2453	0.2708	0.265	0.398	0.2292	0.0698
NLMfusionB [12]	nlm-umd.aronson	a	0.2453	0.2666	0.2541	0.4082	0.2339	0.0693
UniNeHug2 [14]	uneuchatel.savoy	a	0.2439	0.2582	0.264	0.398	0.2308	0.0712
UniGe2 [15]	u.geneva	a	0.2396	0.2705	0.2608	0.3878	0.2361	0.0711
i2r1 [16]	iir.yu	a	0.2391	0.2629	0.2716	0.3898	0.231	0.0668
uta05a [17]	utampere.pirkola	a	0.2385	0.2638	0.2546	0.4163	0.2255	0.0678
i2r2 [16]	iir.yu	a	0.2375	0.2622	0.272	0.3878	0.2296	0.067
UniNeHug2c [14]	uneuchatel.savoy	a	0.2375	0.2662	0.2589	0.3878	0.239	0.0725
uwmtEg05fb	uwaterloo.clarke	a	0.2359	0.2573	0.2552	0.3878	0.2257	0.0712
DUTAdHoc2 [18]	dalianu.yang	m	0.2349	0.2678	0.2725	0.3939	0.2206	0.0648
THUIRgen1S [19]	tsinghua.ma	a	0.2349	0.2663	0.2568	0.4224	0.2214	0.0622
tnog10 [20]	tno.erasmus.kraaij	a	0.2346	0.2607	0.2564	0.3857	0.2227	0.0668
DUTAdHoc1 [18]	dalianu.yang	m	0.2344	0.2718	0.2726	0.402	0.22	0.0645
tnog10p [20]	tno.erasmus.kraaij	a	0.2332	0.2506	0.2555	0.402	0.2173	0.0668
iasl2 [13]	academia.sinica.tsai	a	0.2315	0.2465	0.2487	0.3816	0.2276	0.07
UAmscombGeFb [21]	uamsterdam.aidteam	a	0.2314	0.2638	0.2592	0.4163	0.2271	0.0612
UBIgeneA [22]	suny-buffalo.ruiz	a	0.2262	0.2567	0.2542	0.3633	0.2122	0.0683
OHSUkey [23]	ohsu.hersh	a	0.2233	0.2569	0.2544	0.3735	0.2169	0.0632
NTUgah2 [24]	ntu.chen	a	0.2204	0.2562	0.2498	0.398	0.1996	0.0644
THUIRgen2P [19]	tsinghua.ma	a	0.2177	0.2519	0.2395	0.4143	0.2198	0.0695
NTUgah1 [24]	ntu.chen	a	0.2173	0.2558	0.2513	0.3918	0.1998	0.0615
UniGeNe [15]	u.geneva	a	0.215	0.2364	0.2347	0.3367	0.2237	0.0694
UAmscombGeMl [21]	uamsterdam.aidteam	a	0.2015	0.2325	0.232	0.3551	0.2094	0.0568
uta05i [17]	utampere.pirkola	i	0.198	0.2411	0.229	0.4082	0.2137	0.0547
PDnoSE [25]	upadova.bacchin	a	0.1937	0.2213	0.2183	0.3571	0.2006	0.063
iitprf011003 [26]	iit.urbain	a	0.1913	0.2142	0.2205	0.3612	0.2018	0.065
dcu1 [27]	dublincityu.gurrin	a	0.1851	0.2178	0.2129	0.3816	0.1851	0.0577
dcu2 [27]	dublincityu.gurrin	a	0.1844	0.2234	0.214	0.3959	0.1896	0.0599
SFUshi [28]	simon-fraseru.shi	m	0.1834	0.2072	0.2149	0.3429	0.1898	0.0608
OHSUall [23]	ohsu.hersh	a	0.183	0.2285	0.2221	0.3286	0.1965	0.0592
wim2 [29]	fudan.niu	a	0.1807	0.2006	0.2055	0.3	0.1794	0.057
genome1 [30]	csusm.guillen	a	0.1803	0.2174	0.211	0.3245	0.1749	0.0577
wim1 [29]	fudan.niu	a	0.1781	0.2094	0.2076	0.3347	0.181	0.0592
NCBITHQ [12]	nlm.wilbur	a	0.1777	0.214	0.2192	0.3041	0.1824	0.0526
NCBIMAN [12]	nlm.wilbur	m	0.1747	0.2081	0.2181	0.3122	0.182	0.0519
UICgen1 [31]	uillinois-chicago.liu	a	0.1738	0.2079	0.2046	0.3082	0.1941	0.0579
MARYGEN1 [32]	umaryland.oard	a	0.1729	0.1954	0.1898	0.3041	0.1439	0.0409
PDSESe02 [25]	upadova.bacchin	a	0.1646	0.1928	0.1928	0.3224	0.1904	0.0615
genome2 [30]	csusm.guillen	a	0.1642	0.1931	0.1928	0.298	0.1676	0.0565
UIowa05GN102 [33]	uiowa.eichmann	a	0.1303	0.1861	0.1693	0.2898	0.1671	0.0396
UMD01 [34]	umichigan-dearborn.murphey	a	0.1221	0.1541	0.1435	0.3224	0.1473	0.0321
UIowa05GN101 [33]	uiowa.eichmann	a	0.1095	0.1636	0.1414	0.2857	0.1571	0.026
CCP0 [35]	ucolorado.cohen	m	0.1078	0.1486	0.1311	0.2837	0.1439	0.0203
YAMAHASHI2	utokyo.takahashi	m	0.1022	0.1236	0.1276	0.2653	0.1312	0.0369
YAMAHASHI1	utokyo.takahashi	m	0.1003	0.1224	0.1248	0.2531	0.1267	0.0356
dpsearch2 [36]	datapark.zakharov	m	0.0861	0.1169	0.1034	0.2633	0.1231	0.0278
dpsearch1 [36]	datapark.zakharov	m	0.0827	0.1177	0.1017	0.2551	0.1182	0.0274
asubaral	arizonau.baral	m	0.0797	0.1079	0.0967	0.2714	0.1061	0.0142
CCP1 [35]	ucolorado.cohen	m	0.0554	0.0963	0.0775	0.1878	0.0951	0.0134
UMD02 [34]	umichigan-dearborn.murphey	a	0.0544	0.0703	0.0735	0.1755	0.0843	0.0166
Minimum			0.0544	0.0703	0.0735	0.1755	0.0843	0.0134
Mean			0.1968	0.2258	0.2218	0.3576	0.1976	0.0573
Maximum			0.302	0.3212	0.3155	0.4735	0.2643	0.0766

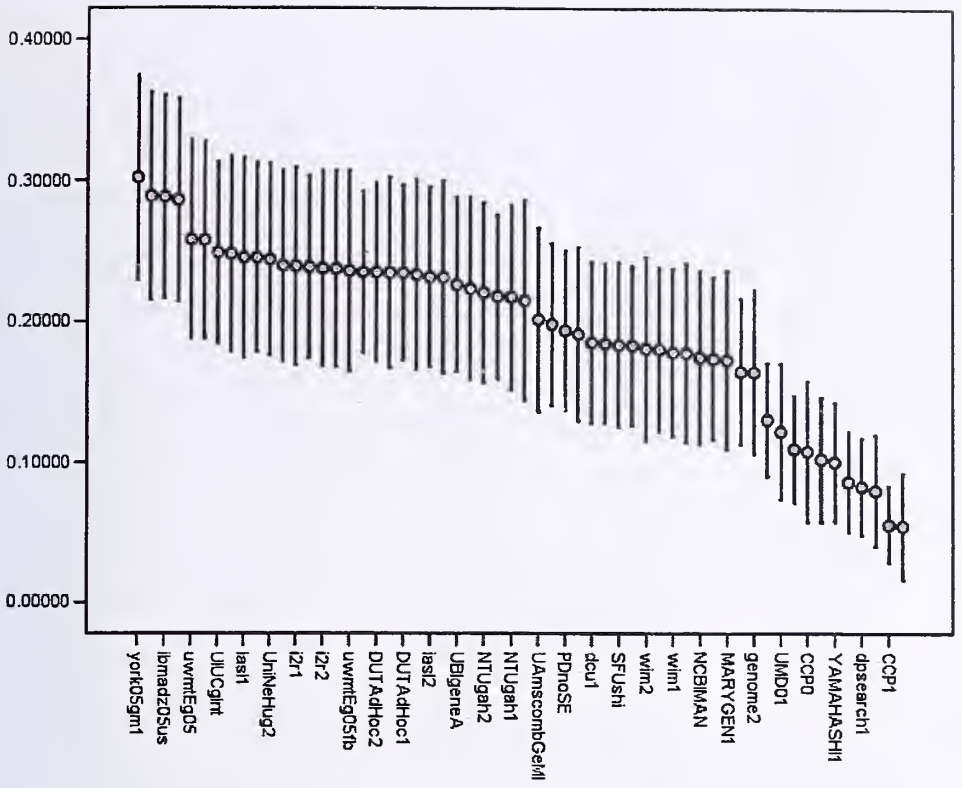

Figure 2 - Run results with 95% confidence intervals, sorted alphabetically.

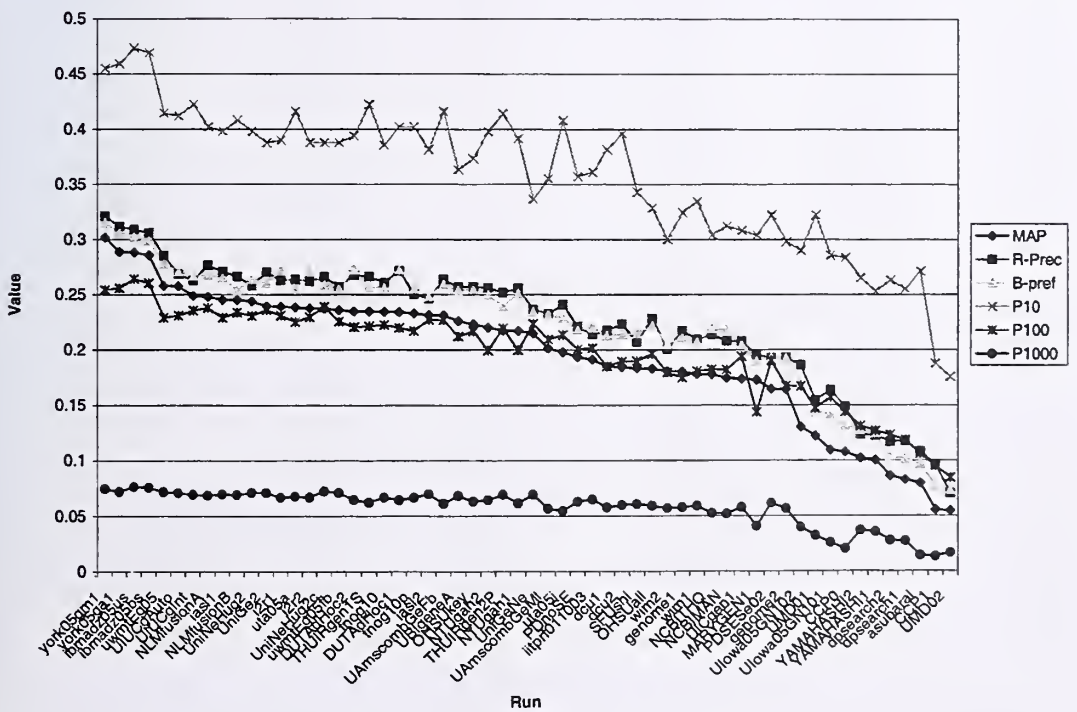

Figure 3 - Run results plotted graphically, sorted by MAP of each run.

Table 5 - Results by topic.

Topic	MAP	R-Prec	B-Pref	P10	P100	P1000
100	0.1691	0.2148	0.1616	0.3569	0.1916	0.0550
101	0.0454	0.0526	0.0285	0.0483	0.0516	0.0141
102	0.0110	0.0172	0.0100	0.0172	0.0091	0.0036
103	0.0603	0.0945	0.0570	0.0948	0.0602	0.0169
104	0.0694	0.0948	0.0582	0.0690	0.0124	0.0023
105	0.1102	0.1703	0.1461	0.4655	0.1586	0.0327
106	0.0625	0.1120	0.1231	0.3138	0.1433	0.0491
107	0.4184	0.4297	0.5289	0.9103	0.5934	0.1373
108	0.1224	0.1973	0.2206	0.4828	0.2788	0.0695
109	0.5347	0.5196	0.6512	0.9190	0.7066	0.1345
110	0.0137	0.0248	0.0154	0.0224	0.0128	0.0055
111	0.2192	0.2985	0.2926	0.3569	0.3140	0.1170
112	0.2508	0.3354	0.2754	0.3586	0.0481	0.0062
113	0.3124	0.3498	0.3164	0.3931	0.0822	0.0096
114	0.3876	0.4364	0.5505	0.8259	0.6697	0.2476
115	0.0378	0.0437	0.0340	0.0534	0.0193	0.0036
116	0.1103	0.1720	0.1456	0.2879	0.1636	0.0359
117	0.3796	0.4739	0.5126	0.8345	0.7409	0.4099
118	0.1343	0.1460	0.1369	0.3276	0.0634	0.0145
119	0.5140	0.5212	0.5075	0.8190	0.3462	0.0493
120	0.5769	0.5421	0.7217	0.9259	0.8091	0.2695
121	0.6205	0.6560	0.6394	0.7983	0.3040	0.0337
122	0.1423	0.2023	0.1590	0.3569	0.1510	0.0320
123	0.0375	0.0708	0.0474	0.1121	0.0493	0.0133
124	0.1519	0.2035	0.1693	0.5103	0.1505	0.0324
125	0.0772	0.0862	0.0708	0.0897	0.0209	0.0028
126	0.1313	0.2172	0.2388	0.3966	0.2979	0.1422
127	0.1015	0.1250	0.0862	0.0759	0.0155	0.0028
128	0.0921	0.1424	0.1062	0.3224	0.1247	0.0366
129	0.0864	0.1393	0.0939	0.1793	0.0984	0.0212
130	0.3390	0.3545	0.3346	0.6362	0.1388	0.0194
131	0.4436	0.4384	0.4230	0.5517	0.2790	0.0343
132	0.1048	0.1558	0.1115	0.2431	0.0966	0.0196
133	0.0328	0.0207	0.0172	0.0172	0.0140	0.0029
134	0.1687	0.1771	0.1582	0.1914	0.0364	0.0069
136	0.0032	0.0000	0.0000	0.0000	0.0019	0.0010
137	0.0676	0.1146	0.0767	0.1776	0.0848	0.0232
138	0.2196	0.2342	0.2029	0.2534	0.0552	0.0089
139	0.3600	0.3941	0.3488	0.5810	0.2052	0.0305
140	0.2700	0.3115	0.2423	0.3810	0.1843	0.0248
141	0.2381	0.2735	0.2053	0.3362	0.2598	0.0699
142	0.4416	0.4608	0.5911	0.8569	0.6409	0.2098
143	0.0031	0.0043	0.0011	0.0034	0.0021	0.0009
144	0.0734	0.0603	0.0431	0.0276	0.0053	0.0009
145	0.3363	0.3761	0.3238	0.5931	0.1852	0.0260
146	0.4808	0.4961	0.6325	0.8466	0.7212	0.3076
147	0.0087	0.0138	0.0057	0.0138	0.0091	0.0040
148	0.0411	0.0376	0.0144	0.0293	0.0407	0.0066
149	0.0286	0.0495	0.0304	0.0603	0.0347	0.0089

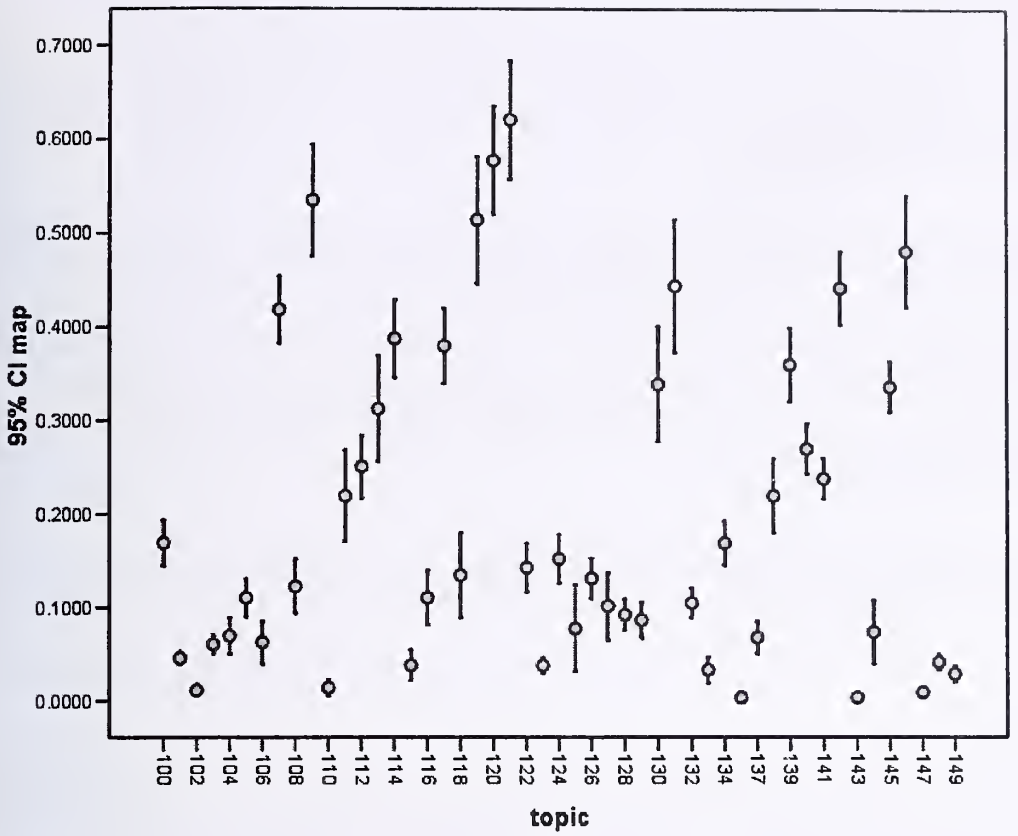

Figure 4 - Results by topic plotted graphically.

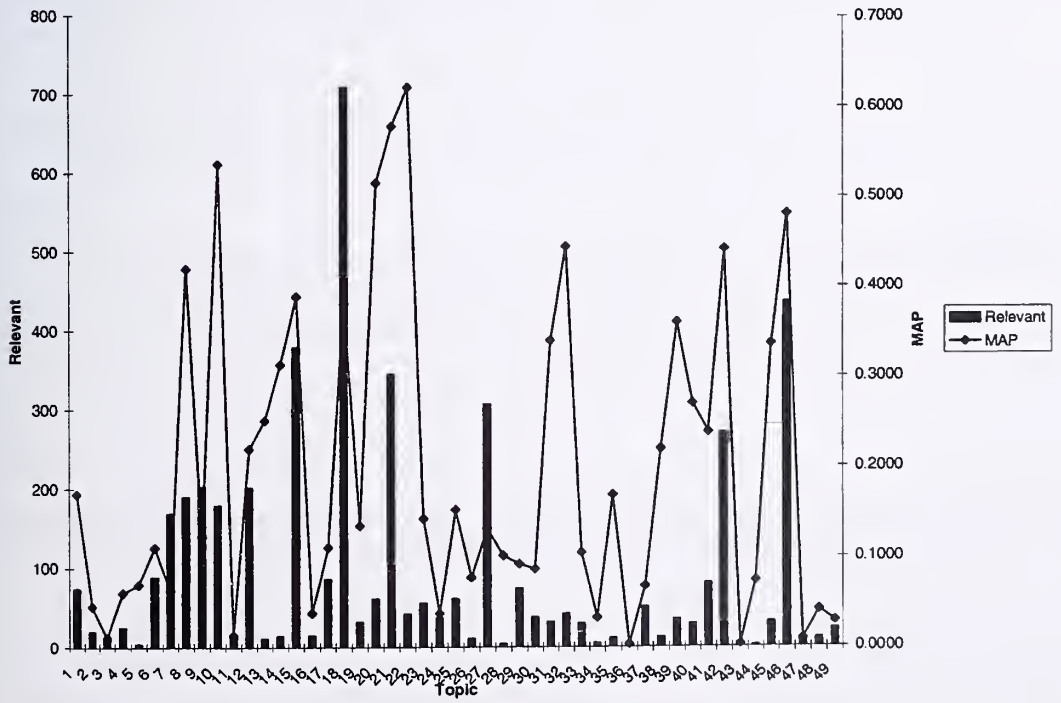

Figure 5 - Comparison of number of relevant documents and MAP for each topic.

Table 6 - Results by generic topic type.

Topics	GTT	MAP	R-Prec	B-Pref	P10	P100	P1000
100-109	Information describing standard methods or protocols for doing some sort of experiment or procedure	0.1603	0.1903	0.1985	0.3678	0.2206	0.0515
110-119	Information describing the role(s) of a gene involved in a disease	0.2360	0.2802	0.2787	0.4279	0.2460	0.0899
120-129	Information describing the role of a gene in a specific biological process	0.2018	0.2385	0.2333	0.3767	0.2021	0.0587
130-139	Information describing interactions (e.g., promote, suppress, inhibit, etc.) between two or more genes in the function of an organ or in a disease	0.1932	0.2099	0.1859	0.2946	0.1013	0.0163
140-149	Information describing one or more mutations of a given gene and its biological impact or role	0.1922	0.2084	0.2090	0.3148	0.2083	0.0659

3. Categorization Task

3.1 Subtasks

The second task for the 2005 track was a full-text document categorization task. It was similar in part to the 2004 categorization task in using data from the Mouse Genome Informatics (MGI, http://www.informatics.jax.org/) system [37] and was a document triage task, where a decision is made on a per-document basis about whether or not to pass a document on for further expert review. It included a repeat of one subtask from last year, the triage of articles for GO annotation [38], and added triage of articles for three other major types of information collected and catalogued by MGI. These include articles about tumor biology [39], embryologic gene expression [40], and alleles of mutant phenotypes [41].

As such, the categorization task assessed how well systems can categorize documents in four separate categories. We used the same utility measure used last year but with different parameters (see below). We created an updated version of the cat_eval program that calculated the utility measure plus recall, precision, and the F score.

3.2 Documents

The documents for the 2005 categorization tasks consisted of the same full-text articles used in 2004. The articles came from three journals over two years, reflecting the full-text data we were able to obtain from Highwire Press: *Journal of Biological Chemistry* (JBC), *Journal of Cell Biology* (JCB), and *Proceedings of the National Academy of Science* (PNAS). These journals have a good proportion of mouse genome articles. Each of the papers from these journals was available in SGML format based on Highwire's document type definition (DTD). Also the same as 2004, we designated articles published in 2002 as training data and those in 2003 as test data.

The documents for the tasks come from a subset of these articles that have the words "mouse" or "mice" or "murine" as described in the 2004 protocol. A crosswalk (look-up) table was provided that matches an identifier for each Highwire article (its file name) to its corresponding PubMed ID (PMID). Table 7 shows the total number of articles and the number in the subset the track used.

The training document collection was 150 megabytes in size compressed and 449 megabytes uncompressed. The test document collection was 140 megabytes compressed and 397 megabytes uncompressed. Many gene names have Greek or other non-English characters, which can present a problem for those attempting to recognize gene names in the text. The Highwire SGML appears to obey the rules posted on the NLM Web site with regards to these characters (http://www.ncbi.nlm.nih.gov/entrez/query/static/entities.html).

3.3 Data

The data for the triage decisions were provided by MGI. They were reformatted in a way to allow easy use by track participants and the cat_eval evaluation program.

3.4 Evaluation Measures

While we again used the utility measure as the primary evaluation measure, we used it in a slightly different way in 2005. This was because there were varying numbers of positive examples for the four different categorization tasks. The framework for evaluation in the categorization task is based on the possibilities in Table 8. The utility measure is often applied in text categorization research and was used by the former TREC Filtering Track. This measure contains coefficients for the utility of retrieving a relevant and retrieving a nonrelevant document. We used a version that was normalized by the best possible score:
$U_{norm} = U_{raw} / U_{max}$

Table 7 - Distribution of documents in training and test sets.

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

Table 8 - Categories for utility measures.

	Relevant (classified)	**Not relevant (not classified)**	**Total**
Retrieved	True positive (TP)	False positive (FP)	All retrieved (AR)
Not retrieved	False negative (FN)	True negative (TN)	All not retrieved (ANR)
	All positive (AP)	All negative (AN)	

For a given test collection of documents to categorize, U_{raw} is calculated as follows:
$U_{raw} = (u_r * TP) + (u_{nr} * FP)$
where:

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

For our purposes, we assume that $u_{nr} = -1$ and solve for u_r assigning MGI's current practice of triaging everything a utility of 0.0:

$0.0 = u_r*AP - AN$
$u_r = AN/AP$

AP and AN are different for each task, as shown in Table 9. (The numbers for GO annotation are slightly different from the 2004 data. This is because additional articles have been triaged by MGI since we used that data last year.)

The u_r values for A and G are fairly close across the training and test collections, while they vary much more for E and especially T. We therefore established a u_r that was the average of that computed for the training and test collections, rounded to the nearest whole number. The resulting values for u_r for each subtask are shown in Table 10. In order to facilitate calculation of the modified version of the utility measure for the 2005 track, we updated the cat_eval program to version 2.0, which included a command-line parameter to set u_r. The training and test data were provided in four files, one for each category (i.e., A, E, G, and T). (The fact that three of those four corresponded to the four nucleotides in DNA was purely coincidental! We could not think of a good way to make a C from embryonic expression.)

Table 9 - Calculating u_r for subtasks.

Subtask	**Training**				**Test**			
	N	AP	AN	u_r	N	AP	AN	u_r
A (alelle)	5837	338	5499	16.27	6043	332	5711	17.20
E (expression)	5837	81	5756	71.06	6043	105	5938	56.55
G (GO annotation)	5837	462	5375	11.63	6043	518	5525	10.67
T (tumor)	5837	36	5801	161.14	6043	20	6023	301.15

Table 10 - Values of u_r for subtasks.

Subtask	u_r
A (alelle)	17
E (expression)	64
G (GO annotation)	11
T (tumor)	231

A common question that emerged was, what resources can be legitimately used to aid in categorizing the documents? In general, groups could use anything, including resources on the MGI Web site. The only resource they could not use was the direct data itself, i.e., data that was directly linked to the PMID or the associated MGI unique identifier. Thus, they could not go into the MGI database (or any other aggregated resource such as Entrez Gene or SOURCE) and pull out GO codes, tumor terms, mutant phenotypes, or any other data that was explicitly linked to a document. But anything else was fair game.

3.5 Results

A total of 46-48 runs were submitted for each of the four tasks. The results varied widely by subtask. The highest results were obtained in the tumor subtask, followed by the allele and expression subtasks very close to each other, and the GO subtask substantially lower. In light of the concern about the GO subtask and the inability of any feature beyond the MeSH term *Mice* to improve performance in 2004, this year's results are reassuring that document triage can potentially be helpful to model organism database curators. Table 11 shows the best and median U_{norm} values. Tables 12-15 show the results of the four subtasks; Figures 6-9 depict these results graphically.

From these results, it is clear that the GO task is somewhat different than the other tasks. The best utility scores that participants were able to achieve were in the 0.50-0.60 range, which were much lower than for the other three tasks. Another interesting observation is the u_r factor for the best performing task, tumor biology at 231, was the highest among the tasks, while the lowest occurred for the worst performing task, GO, at 11. While a high u_r leads to an increasing preference for high recall over precision, a u_r of 11 is still substantial compared to typical, more balanced classification tasks where the goal is often to optimize F-measure. Further investigation is needed to understand why the GO task appears more difficult than the other three. A separate analysis of the similar 2004 data shows that the individual GO codes are very sparsely represented in the training and test collections. This observation combined with assuming that correctly categorizing a paper is highly dependent upon the specific GO codes associated with the paper may explain why the GO task is more heterogeneous and therefore complex than the other tasks [6].

Table 11 - Best and median results for each subtask.

Subtask	Best u_{norm}	Median u_{norm}	u_r
A (alelle)	0.871	0.7773	17
E (expression)	0.8711	0.6413	64
G (GO annotation)	0.587	0.4575	11
T (tumor)	0.9433	0.761	231

4. Future Directions

The TREC Genomics 2005 Genomics Track was again carried out with much participation and enthusiasm. To prepare for the 2006 track, we created an on-line survey for members of the track email list. A total of 26 people responded to the survey, the results of which can be found at http://ir.ohsu.edu/genomics/2005survey.html. In summary, the results indicate that there is a strong desire for full-text journal articles for the ad hoc task and an information extraction task as the second task for the track in 2006.

Acknowledgements

The TREC Genomics Track is funded by grant ITR-0325160 from the U.S. National Science Foundation. The following individuals carried out interviews with biologists to obtain topics for the ad hoc task: Marti Hearst, Laura Ross, Leonie IJzereef, Dina Demner, Sharon Yang, Phoebe Roberts, William Cohen, Kevin Cohen, Paul Thompson, LV Subramaniam, and Jay Urbain. The track also thanks Ellen Voorhees, Ian Soboroff, and Lori Buckland of NIST for their help in various ways.

Table 12 - Results of allele subtask by run, sorted by utility measure.

Tag	Group	Precision	Recall	F-Score	Utility
aibmadz05s [10]	ibm.zhang	0.4669	0.9337	0.6225	0.871
ABBR003SThr [42]	ibm.kanungo	0.4062	0.9458	0.5683	0.8645
ABBR003 [42]	ibm.kanungo	0.3686	0.9548	0.5319	0.8586
aibmadz05m1 [10]	ibm.zhang	0.5076	0.9006	0.6493	0.8492
aibmadz05m2 [10]	ibm.zhang	0.5025	0.9006	0.6451	0.8482
cuhkrun3A [43]	cuhk.lam	0.3442	0.9548	0.506	0.8478
THUIRgenA1p1 [19]	tsinghua.ma	0.4902	0.9006	0.6348	0.8455
cuhkrun2A [43]	cuhk.lam	0.3316	0.9578	0.4926	0.8443
aFduMarsII [29]	fudan.niu	0.4195	0.9187	0.576	0.8439
aNTUMAC [24]	ntu.chen	0.3439	0.9488	0.5048	0.8423
aFduMarsI [29]	fudan.niu	0.4754	0.9006	0.6223	0.8421
ASVMN03 [42]	ibm.kanungo	0.4019	0.9127	0.558	0.8327
aNLMB [12]	nlm-umd.aronson	0.3391	0.9398	0.4984	0.832
aDIMACSl9w [44]	rutgers.dayanik	0.4357	0.8976	0.5866	0.8292
THUIRgA0p9x [19]	tsinghua.ma	0.5414	0.8675	0.6667	0.8242
cuhkrun1 [43]	cuhk.lam	0.3257	0.9367	0.4833	0.8226
aDIMACSg9md [44]	rutgers.dayanik	0.4509	0.8855	0.5976	0.8221
aDIMACSl9md [44]	rutgers.dayanik	0.3844	0.9066	0.5399	0.8212
aDIMACSg9w [44]	rutgers.dayanik	0.4882	0.8705	0.6255	0.8168
NLM2A [12]	nlm-umd.aronson	0.4332	0.8795	0.5805	0.8118
AOHSUVP [23]	ohsu.hersh	0.3556	0.8976	0.5094	0.8019
aFduMarsIII [29]	fudan.niu	0.3254	0.9096	0.4794	0.7987
aDUTCat1 [18]	dalianu.yang	0.2858	0.9307	0.4374	0.7939
AOHSUSL [23]	ohsu.hersh	0.3448	0.8765	0.4949	0.7785
aQUT14 [45]	queensu.shatkay	0.3582	0.8675	0.507	0.776
AOHSUBF [23]	ohsu.hersh	0.3007	0.8976	0.4505	0.7748
aIBMIRLrul [46]	ibm-india.ramakrishnan	0.3185	0.8855	0.4685	0.7741
Ameta [47]	uwisconsin.craven	0.3031	0.8946	0.4527	0.7736
Apars [47]	uwisconsin.craven	0.2601	0.9277	0.4063	0.7725
aIBMIRLsvm [46]	ibm-india.ramakrishnan	0.2982	0.8946	0.4473	0.7707
aDUTCat2 [18]	dalianu.yang	0.262	0.9217	0.408	0.769
aMUSCUIUC3 [11]	uiuc.zhai	0.4281	0.8072	0.5595	0.7438
Afull [47]	uwisconsin.craven	0.2718	0.8825	0.4156	0.7434
aMUSCUIUC2 [11]	uiuc.zhai	0.5501	0.7771	0.6442	0.7397
aQUNB8 [45]	queensu.shatkay	0.3182	0.8464	0.4626	0.7397
aIBMIRLmet [46]	ibm-india.ramakrishnan	0.32	0.8434	0.464	0.738
ABPLUS [20]	erasmus.kors	0.241	0.8916	0.3795	0.7264
aUCHSCnb1En3 [35]	ucolorado.cohen	0.508	0.7651	0.6106	0.7215
aQUT11 [45]	queensu.shatkay	0.3785	0.7741	0.5084	0.6993
aUCHSCnb1En4 [35]	ucolorado.cohen	0.6091	0.6476	0.6277	0.6231
aMUSCUIUC1 [11]	uiuc.zhai	0.6678	0.6054	0.6351	0.5877
aUCHSCsvm [35]	ucolorado.cohen	0.7957	0.4458	0.5714	0.4391
aNLMF [12]	nlm-umd.aronson	0.2219	0.5301	0.3129	0.4208
LPC6	langpower.yang	0.4281	0.4307	0.4294	0.3969
FTA [20]	erasmus.kors	0.3562	0.3916	0.373	0.3499
aLRIk1	uparis-sud.kodratoff	0.2331	0.259	0.2454	0.2089
aLRIk3	uparis-sud.kodratoff	0.2191	0.262	0.2387	0.2071
aLRIk2	uparis-sud.kodratoff	0.2306	0.25	0.2399	0.2009
Minimum		0.2191	0.25	0.2387	0.2009
Median		0.3572	0.8931	0.5065	0.77725
Maximum		0.7957	0.9578	0.6667	0.871

Table 13 - Results of expression subtask by run, sorted by utility measure.

Tag	Group	Precision	Recall	F-Score	Utility
eFduMarsI [29]	fudan.niu	0.1899	0.9333	0.3156	0.8711
eFduMarsII [29]	fudan.niu	0.1899	0.9333	0.3156	0.8711
eDUTCat1 [18]	dalianu.yang	0.1364	0.9429	0.2383	0.8496
eDIMACSl9w [44]	rutgers.dayanik	0.2026	0.9048	0.331	0.8491
eibmadz05s [10]	ibm.zhang	0.1437	0.9333	0.249	0.8464
eibmadz05m2 [10]	ibm.zhang	0.2109	0.8857	0.3407	0.8339
cuhkrun2E [43]	cuhk.lam	0.126	0.9333	0.222	0.8321
cuhkrun3E [43]	cuhk.lam	0.1481	0.9143	0.255	0.8321
EBBR0006SThr [42]	ibm.kanungo	0.1228	0.9333	0.2171	0.8292
THUIRgenE1p8 [19]	tsinghua.ma	0.1322	0.9238	0.2312	0.829
eibmadz05m1 [10]	ibm.zhang	0.2201	0.8762	0.3518	0.8277
EBBR0006 [42]	ibm.kanungo	0.1211	0.9333	0.2144	0.8275
eDUTCat2 [18]	dalianu.yang	0.1104	0.9429	0.1976	0.8241
ESVMN075 [42]	ibm.kanungo	0.1265	0.9143	0.2222	0.8156
cuhkrun1E [43]	cuhk.lam	0.1119	0.9143	0.1994	0.8009
eDIMACSg9w [44]	rutgers.dayanik	0.2444	0.8381	0.3785	0.7976
eFduMarsIII [29]	fudan.niu	0.0794	0.9524	0.1466	0.7799
eNTUMAC [24]	ntu.chen	0.1593	0.819	0.2667	0.7515
Epars [47]	uwisconsin.craven	0.0818	0.8857	0.1498	0.7304
eIBMIRLsvm [46]	ibm-india.ramakrishnan	0.0571	0.9238	0.1075	0.6854
ABPLUSE [20]	erasmus.kors	0.0841	0.819	0.1525	0.6796
eDIMACSg9md [44]	rutgers.dayanik	0.1575	0.7333	0.2593	0.672
Emeta [47]	uwisconsin.craven	0.1273	0.7333	0.2169	0.6548
eDIMACSl9md [44]	rutgers.dayanik	0.1054	0.7238	0.184	0.6278
NLM2E [12]	nlm-umd.aronson	0.2863	0.6381	0.3953	0.6132
EOHSUBF [23]	ohsu.hersh	0.0405	0.9619	0.0777	0.6058
Efull [47]	uwisconsin.craven	0.0636	0.781	0.1176	0.6012
EOHSUVP [23]	ohsu.hersh	0.0693	0.7429	0.1267	0.5869
EOHSUSL [23]	ohsu.hersh	0.0365	0.9905	0.0705	0.5824
eIBMIRLmet [46]	ibm-india.ramakrishnan	0.0627	0.7333	0.1155	0.5621
eIBMIRLrul [46]	ibm-india.ramakrishnan	0.0642	0.7238	0.1179	0.5589
eQUNB11 [45]	queensu.shatkay	0.1086	0.6381	0.1856	0.5563
eQUT18 [45]	queensu.shatkay	0.0967	0.5238	0.1632	0.4473
eMUSCUIUC1 [11]	uiuc.zhai	0.2269	0.4667	0.3053	0.4418
eMUSCUIUC3 [11]	uiuc.zhai	0.1572	0.4762	0.2364	0.4363
eQUNB19 [45]	queensu.shatkay	0.1132	0.4571	0.1815	0.4012
eUCHSCnb1En4 [35]	ucolorado.cohen	0.52	0.3714	0.4333	0.3661
FTE [20]	erasmus.kors	0.0835	0.4095	0.1387	0.3393
eUCHSCnb1En3 [35]	ucolorado.cohen	0.5714	0.3429	0.4286	0.3388
eNLMF [12]	nlm-umd.aronson	0.129	0.2286	0.1649	0.2045
eNLMKNN [12]	nlm-umd.aronson	0.0519	0.2381	0.0852	0.1701
eLRIk3	uparis-sud.kodratoff	0.0828	0.1238	0.0992	0.1024
eLRIk1	uparis-sud.kodratoff	0.1026	0.1143	0.1081	0.0987
eLRIk2	uparis-sud.kodratoff	0.1026	0.1143	0.1081	0.0987
eUCHSCsvm [35]	ucolorado.cohen	1	0.0381	0.0734	0.0381
eMUSCUIUC2 [11]	uiuc.zhai	0	0	0	-0.0074
Minimum		0	0	0	-0.0074
Median		0.12195	0.8	0.1985	0.6413
Maximum		1	0.9905	0.4333	0.8711

Table 14 - Results of GO subtask by run, sorted by utility measure.

Tag	Group	Precision	Recall	F-Score	Utility
gFduMarsII [29]	fudan.niu	0.2122	0.8861	0.3424	0.587
gFduMarsI [29]	fudan.niu	0.2644	0.778	0.3947	0.5813
gIBMIRLmet [46]	ibm-india.ramakrishnan	0.2028	0.9015	0.3311	0.5793
gDUTCat1 [18]	dalianu.yang	0.1914	0.9286	0.3174	0.572
gIBMIRLrul [46]	ibm-india.ramakrishnan	0.1883	0.9286	0.3132	0.5648
gIBMIRLsvm [46]	ibm-india.ramakrishnan	0.2069	0.8668	0.3341	0.5648
gFduMarsIII [29]	fudan.niu	0.191	0.9093	0.3157	0.5591
GBBR004 [42]	ibm.kanungo	0.1947	0.8938	0.3198	0.5577
GOHSUBF [23]	ohsu.hersh	0.1889	0.9093	0.3127	0.5542
GAbsBBR0083 [42]	ibm.kanungo	0.2524	0.7548	0.3783	0.5516
GOHSUVP [23]	ohsu.hersh	0.2308	0.7819	0.3564	0.5449
GSVMN08 [42]	ibm.kanungo	0.2038	0.8436	0.3283	0.5441
gDUTCat2 [18]	dalianu.yang	0.1779	0.9363	0.2989	0.5428
gNTUMAC [24]	ntu.chen	0.1873	0.8803	0.3089	0.5332
gibmadz05m2 [10]	ibm.zhang	0.3179	0.6216	0.4206	0.5004
gibmadz05m1 [10]	ibm.zhang	0.3216	0.6178	0.423	0.4993
ABPLUSG [20]	erasmus.kors	0.2178	0.7259	0.3351	0.4889
gDIMACSl9w [44]	rutgers.dayanik	0.245	0.668	0.3585	0.4809
gibmadz05s [10]	ibm.zhang	0.3226	0.583	0.4154	0.4717
GOHSUSL [23]	ohsu.hersh	0.2536	0.6429	0.3637	0.4709
gDIMACSl9md [44]	rutgers.dayanik	0.2425	0.6564	0.3542	0.47
cuhkrun1G [43]	cuhk.lam	0.2706	0.6139	0.3757	0.4635
gDIMACSg9md [44]	rutgers.dayanik	0.2529	0.6293	0.3608	0.4603
NLM2G [12]	nlm-umd.aronson	0.3223	0.5656	0.4107	0.4575
Gpars [47]	uwisconsin.craven	0.1862	0.7587	0.299	0.4572
gDIMACSg9w [44]	rutgers.dayanik	0.2754	0.5965	0.3768	0.4538
THUIRgenGMNG [19]	tsinghua.ma	0.2107	0.6776	0.3214	0.4468
Gmeta [47]	uwisconsin.craven	0.1689	0.7934	0.2785	0.4386
NLM1G [12]	nlm-umd.aronson	0.316	0.5405	0.3989	0.4342
cuhkrun2G [43]	cuhk.lam	0.2109	0.6506	0.3185	0.4293
Gfull [47]	uwisconsin.craven	0.1904	0.6988	0.2993	0.4287
THUIRgenG1p1 [19]	tsinghua.ma	0.1827	0.6506	0.2852	0.3859
gQUNB15 [45]	queensu.shatkay	0.2102	0.5676	0.3067	0.3736
gNLMF [12]	nlm-umd.aronson	0.1887	0.6062	0.2878	0.3693
gQUT22 [45]	queensu.shatkay	0.1811	0.6158	0.2799	0.3628
gQUNB12 [45]	queensu.shatkay	0.1603	0.6602	0.258	0.3459
cuhkrun3G [43]	cuhk.lam	0.1651	0.5637	0.2554	0.3045
gUCHSCnb1En3 [35]	ucolorado.cohen	0.4234	0.3417	0.3782	0.2994
gMUSCUIUC1 [11]	uiuc.zhai	0.393	0.2799	0.3269	0.2406
FTG [20]	erasmus.kors	0.2211	0.2876	0.25	0.1955
gUCHSCnb1En4 [35]	ucolorado.cohen	0.5542	0.1776	0.269	0.1646
gUCHSCsvm [35]	ucolorado.cohen	0.406	0.1834	0.2527	0.159
gMUSCUIUC3 [11]	uiuc.zhai	0.0891	0.3456	0.1416	0.0242
gLRIk3	uparis-sud.kodratoff	0.0998	0.1158	0.1072	0.0209
gLRIk2	uparis-sud.kodratoff	0.1	0.1023	0.1011	0.0186
gLRIk1	uparis-sud.kodratoff	0.0938	0.1023	0.0979	0.0125
gMUSCUIUC2 [11]	uiuc.zhai	0.0706	0.1737	0.1004	-0.0342
Minimum		0.0706	0.1023	0.0979	-0.0342
Median		0.2102	0.6506	0.3185	0.4575
Maximum		0.5542	0.9363	0.423	0.587

Table 15 - Results of tumor subtask by run, sorted by utility measure.

Tag	Group	Precision	Recall	F-Score	Utility
tDIMACSg9w [44]	rutgers.dayanik	0.0709	1	0.1325	0.9433
TSVM0035 [42]	ibm.kanungo	0.0685	1	0.1282	0.9411
tDIMACSg9md [44]	rutgers.dayanik	0.0556	1	0.1053	0.9264
tFduMarsI [29]	fudan.niu	0.1061	0.95	0.191	0.9154
tFduMarsII [29]	fudan.niu	0.099	0.95	0.1792	0.9126
tIBMIRLmet [46]	ibm-india.ramakrishnan	0.0945	0.95	0.1719	0.9106
tDIMACSl9w [44]	rutgers.dayanik	0.0444	1	0.0851	0.9069
TBBR0004SThr [42]	ibm.kanungo	0.0436	1	0.0835	0.905
cuhkrun3T [43]	cuhk.lam	0.0426	1	0.0818	0.9028
tibmadz05m2 [10]	ibm.zhang	0.0757	0.95	0.1402	0.8998
tDUTCat1 [18]	dalianu.yang	0.0745	0.95	0.1382	0.8989
tibmadz05s [10]	ibm.zhang	0.0688	0.95	0.1284	0.8944
tibmadz05m1 [10]	ibm.zhang	0.0674	0.95	0.1258	0.8931
TBBR0004 [42]	ibm.kanungo	0.0376	1	0.0725	0.8892
tDUTCat2 [18]	dalianu.yang	0.035	1	0.0677	0.8807
tNTUMACwj [24]	ntu.chen	0.0518	0.95	0.0982	0.8747
tIBMIRLrul [46]	ibm-india.ramakrishnan	0.0415	0.95	0.0795	0.855
cuhkrun1T [43]	cuhk.lam	0.0769	0.9	0.1417	0.8532
tFduMarsIII [29]	fudan.niu	0.0286	1	0.0556	0.8528
tNTUMAC [24]	ntu.chen	0.0526	0.9	0.0994	0.8299
tDIMACSl9md [44]	rutgers.dayanik	0.0323	0.95	0.0625	0.8268
Tpars [47]	uwisconsin.craven	0.0317	0.95	0.0613	0.8242
ABPLUST [20]	erasmus.kors	0.0314	0.95	0.0607	0.8229
Tfull [47]	uwisconsin.craven	0.0443	0.9	0.0845	0.816
Tmeta [47]	uwisconsin.craven	0.0523	0.85	0.0986	0.7833
THUIRgenT1p5 [19]	tsinghua.ma	0.0213	0.95	0.0417	0.761
TOHSUSL [23]	ohsu.hersh	0.0254	0.9	0.0493	0.7502
tQUNB3 [45]	queensu.shatkay	0.0244	0.9	0.0474	0.7439
TOHSUBF [23]	ohsu.hersh	0.0192	0.95	0.0376	0.7396
TOHSUVP [23]	ohsu.hersh	0.0237	0.9	0.0462	0.7394
tMUSCUIUC3 [11]	uiuc.zhai	0.3182	0.7	0.4375	0.6935
tIBMIRLsvm [46]	ibm-india.ramakrishnan	0.0308	0.8	0.0593	0.6909
tQUT10 [45]	queensu.shatkay	0.0132	1	0.026	0.6758
tMUSCUIUC2 [11]	uiuc.zhai	0.0828	0.7	0.1481	0.6665
tQUT14 [45]	queensu.shatkay	0.3095	0.65	0.4194	0.6437
NLM1T [12]	nlm-umd.aronson	0.0813	0.65	0.1444	0.6182
NLM2T [12]	nlm-umd.aronson	0.0813	0.65	0.1444	0.6182
tMUSCUIUC1 [11]	uiuc.zhai	0.3429	0.6	0.4364	0.595
tNTUMACasem [24]	ntu.chen	0.0339	0.65	0.0645	0.5699
LPC7	langpower.yang	0.3548	0.55	0.4314	0.5457
FTT [20]	erasmus.kors	0.0893	0.5	0.1515	0.4779
tNLMF [12]	nlm-umd.aronson	0.0207	0.55	0.0399	0.4372
cuhkrun2T [43]	cuhk.lam	0.0268	0.4	0.0503	0.3372
tUCHSCnb1En3 [35]	ucolorado.cohen	0.1935	0.3	0.2353	0.2946
tUCHSCnb1En4 [35]	ucolorado.cohen	0.375	0.15	0.2143	0.1489
tLRIk2	uparis-sud.kodratoff	0.0909	0.1	0.0952	0.0957
tLRIk1	uparis-sud.kodratoff	0.087	0.1	0.093	0.0955
tLRIk3	uparis-sud.kodratoff	0.069	0.1	0.0816	0.0942
tUCHSCsvm [35]	ucolorado.cohen	1	0.05	0.0952	0.05
Tcsusm2 [30]	csusm.guillen	0.0256	0.05	0.0339	0.0418
Tcsusm1 [30]	csusm.guillen	0.0244	0.05	0.0328	0.0413
Minimum		0.0132	0.05	0.026	0.0413
Median		0.0526	0.9	0.0952	0.761
Max		1	1	0.4375	0.9433

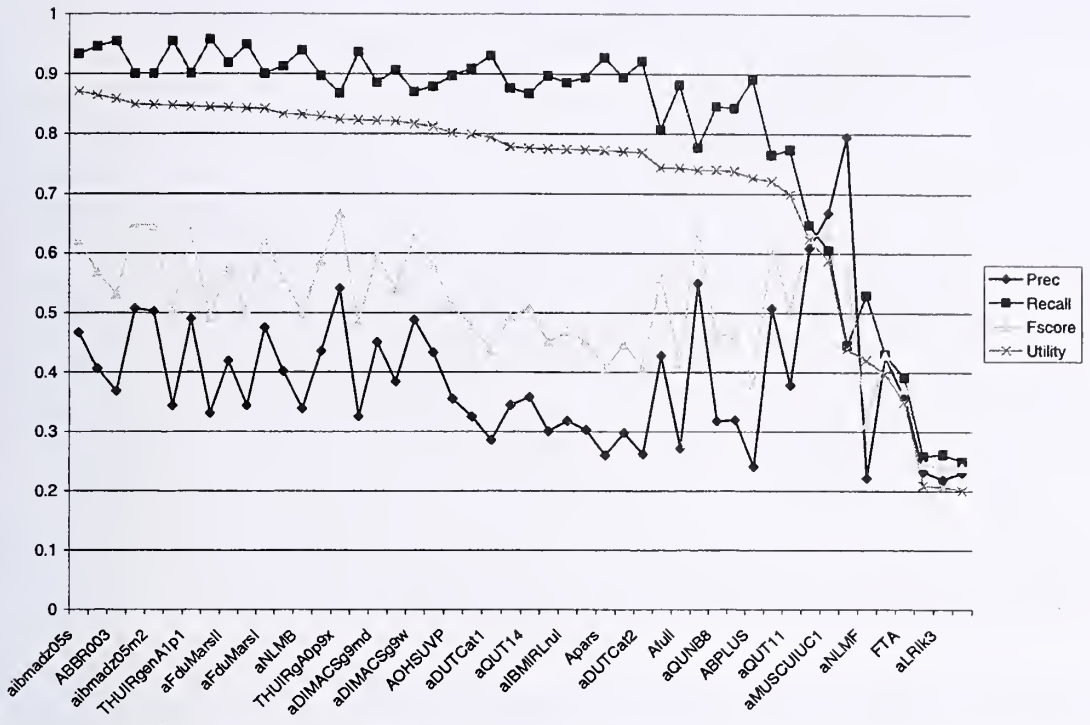

Figure 6 - Results of allele subtask by run displayed graphically, sorted by utility measure.

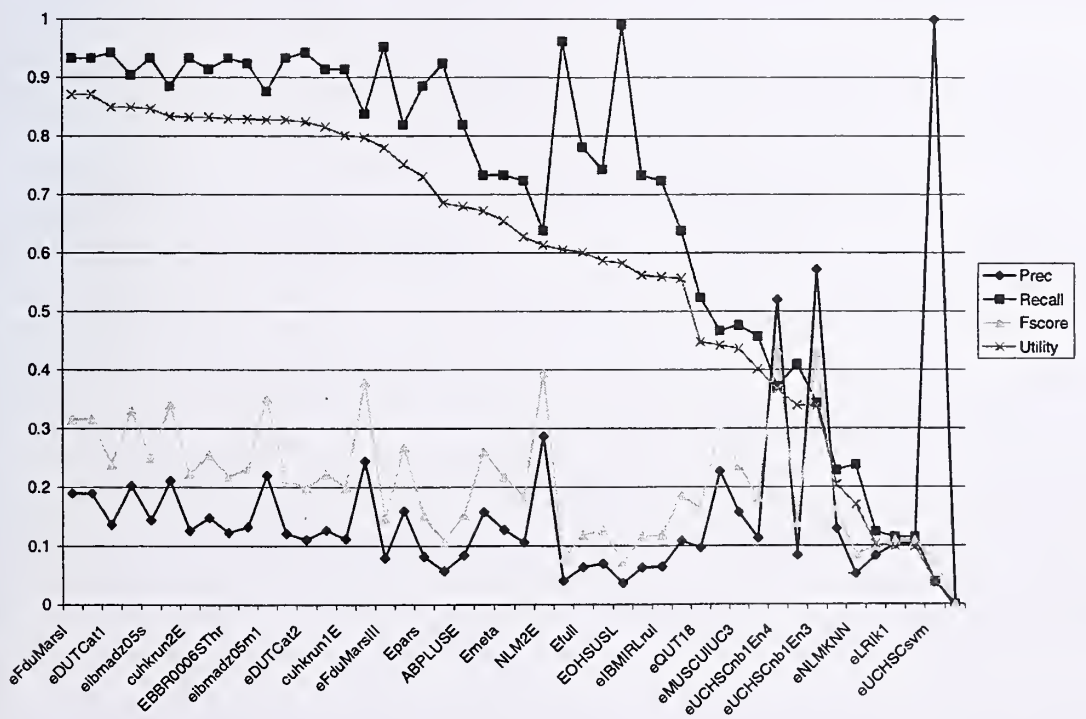

Figure 7 - Results of expression subtask by run displayed graphically, sorted by utility measure.

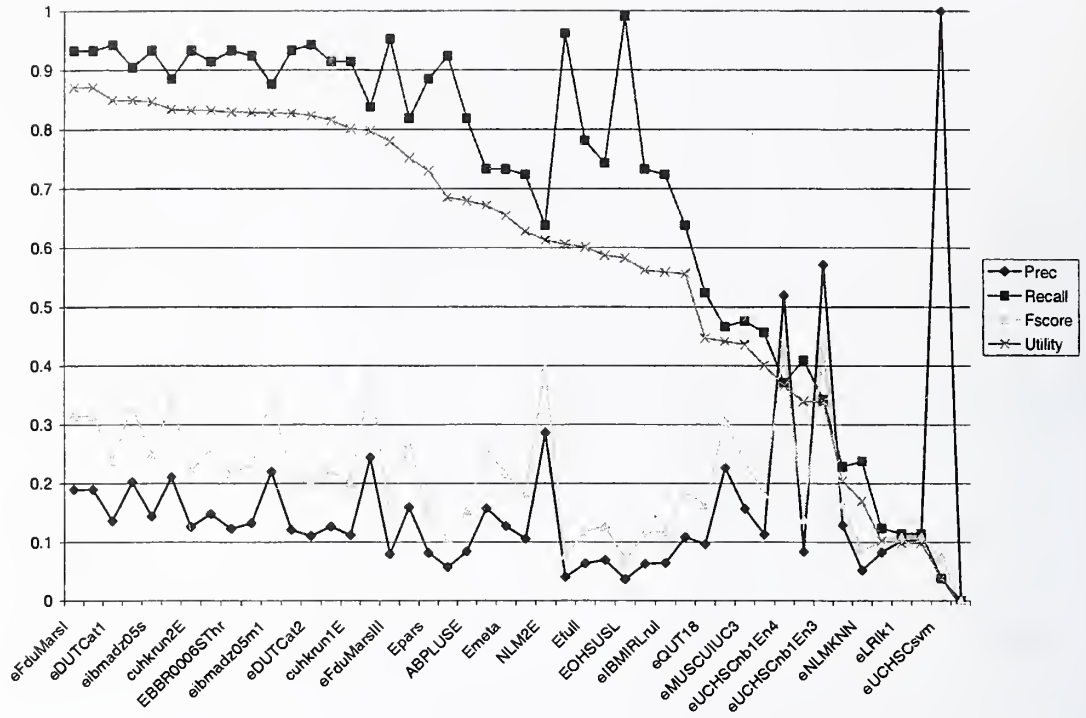

Figure 8 - Results of GO subtask by run displayed graphically, sorted by utility measure.

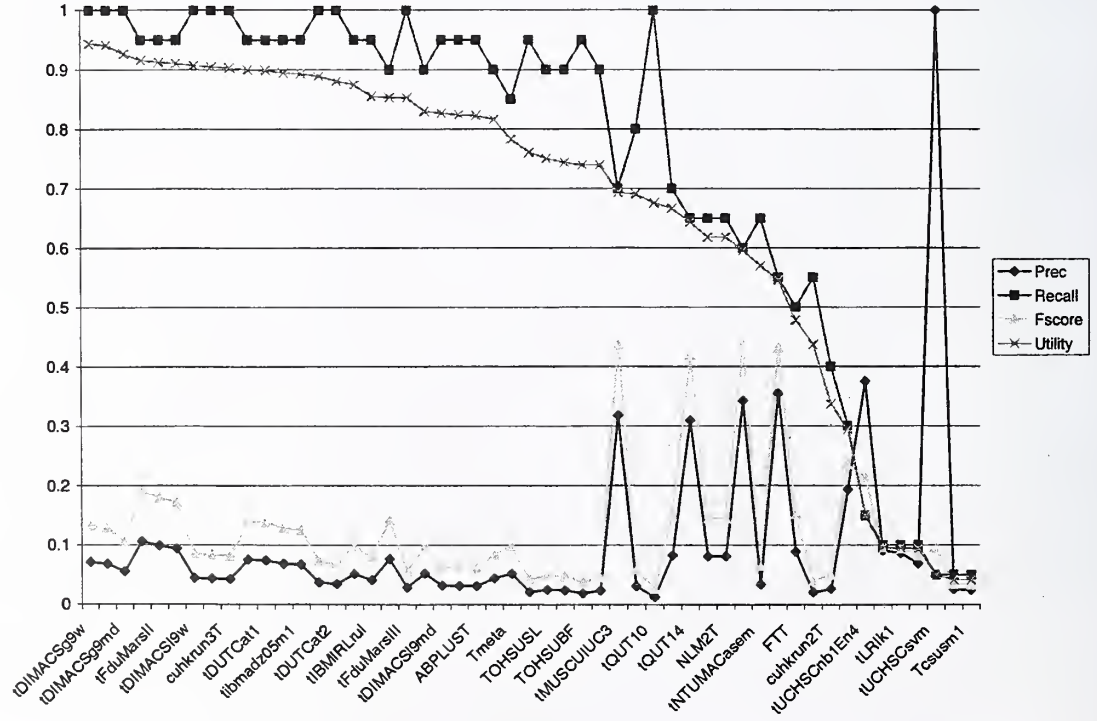

Figure 9 - Results of tumor subtask by run displayed graphically, sorted by utility measure.

References

1. Mobasheri A, et al., Post-genomic applications of tissue microarrays: basic research, prognostic oncology, clinical genomics and drug discovery. Histology and Histopathology, 2004. 19: 325-335.
2. Hirschman L, et al., Accomplishments and challenges in literature data mining for biology. Bioinformatics, 2002. 18: 1553-1561.
3. Cohen AM and Hersh WR, A survey of current work in biomedical text mining. Briefings in Bioinformatics, 2005. 6: 57-71.
4. Hersh WR and Bhupatiraju RT. TREC genomics track overview. The Twelfth Text Retrieval Conference (TREC 2003). 2003. Gaithersburg, MD: NIST. 14-23. http://trec.nist.gov/pubs/trec12/papers/GENOMICS.OVERVIEW3.pdf.
5. Hersh WR, et al., Enhancing access to the bibliome: the TREC 2004 Genomics Track. Journal of Biomedical Discovery and Collaboration, 2006: in press.
6. Cohen AM and Hersh WR, The TREC 2004 Genomics Track categorization task: classifying full-text biomedical documents. Journal of Biomedical Discovery and Collaboration, 2006: in press.
7. Cohen AM, et al., Reducing workload in systematic review preparation using automated citation classification. Journal of the American Medical Informatics Association, 2006: in press.
8. Buckley C and Voorhees EM. Retrieval evaluation with incomplete information. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2004. Sheffield, England: ACM Press. 25-32.
9. Huang X, Zhong M, and Si L. York University at TREC 2005: Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/yorku-huang2.geo.pdf.
10. Ando RK, Dredze M, and Zhang T. TREC 2005 Genomics Track experiments at IBM Watson. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/ibm-tjwatson.geo.pdf.
11. Zhai C, et al. UIUC/MUSC at TREC 2005 Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/uillinois-uc.geo.pdf.
12. Aronson AR, et al. Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/nlm-umd.geo.pdf.
13. Tsai TH, et al. Enhance genomic IR with term variation and expansion: experience of the IASL Group at Genomic Track 2005. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/academia-sinica.geo.pdf.
14. Abdou S, Savoy J, and Ruch P. Evaluation of stemming, query expansion and manual indexing approaches for the genomic task. The Fourteenth Text REtrieval Conference

Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/uneuchatel.geo.pdf.

15. Ruch P, et al. Report on the TREC 2005 experiment: Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/uhospital-geneva.geo.pdf.
16. Yu N, et al. TREC 2005 Genomics Track at I2R. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/inst-infocomm.geo.pdf.
17. Pirkola A. TREC 2005 Genomics Track experiments at UTA. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/utampere.geo.pdf.
18. Yang Z, et al. TREC 2005 Genomics Track experiments at DUTAI. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/dalianu.geo.pdf.
19. Li J, et al. Learning domain-specific knowledge from context - THUIR at TREC 2005 Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/tsinghuau-ma.geo.pdf.
20. Schijvenaars BJA, et al. TREC 2005 Genomics Track - a concept-based approach to text categorization. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/erasmus-tno.geo.pdf.
21. Meij E, et al. Combining thesauri-based methods for biomedical retrieval. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/uamsterdam-infoinst.geo.pdf.
22. Ruiz ME and Southwick SB. UB at CLEF 2005: bilingual CLIR and medical image retrieval tasks. 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Springer Lecture Notes in Computer Science. 2005. Vienna, Austria: Springer-Verlag. in press.

23. Cohen AM, Yang J, and Hersh WR. A comparison of techniques for classification and ad hoc retrieval of biomedical documents. The Fourteenth Text Retrieval Conference - TREC 2005. 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/ohsu-geo.pdf.
24. Lin KHY, Hou WJ, and Chen HH. Retrieval of biomedical documents by prioritizing key phrases. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/ntu.geo.adhoc.pdf.
25. Bacchin M and Melucci M. Symbol-based query expansion experiments at TREC 2005 Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/upadova.geo.pdf.

26. Urbain J, Goharian N, and Frieder O. IIT TREC 2005: Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/iit-urbain.geo.pdf.
27. Camous F, et al. Structural term extraction for expansion of template-based genomic queries. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/dublincu.geo.pdf.
28. Shi Z, et al. Synonym-based expansion and boosting-based re-ranking: a two-phase approach for genomic information retrieval. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/simon-fraseru.geo.pdf.
29. Niu J, et al. WIM at TREC 2005. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/fudanu-sun.geo.ent.pdf.
30. Guillen R. CSUSM at TREC 2005: Genomics and Enterprise Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/calstateu-sanmarcos.geo.ent.pdf.
31. Zhou W and Yu C. Experiment report of TREC 2005 Genomics Track ad hoc retrieval task. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/uillinois-chicago.geo.pdf.
32. Lin J, et al. A menagerie of tracks at Maryland: HARD, Enterprise, QA, and Genomics, oh my! The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/umaryland-lin.hard.ent.qa.geo.pdf.
33. Eichmann D and Srinivasan P. Experiments in questions and relationships at the University of Iowa. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/uiowa.geo.qa.pdf.
34. Huang L, Chen Z, and Murphey YL. UM-D at TREC 2005: Genomics Track. The Fourteenth Text Retrieval Conference - TREC 2005. 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/umich-dearborn.geo.pdf.
35. Caporaso JG, et al. Concept recognition and the TREC genomics tasks. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/ucolorado-hsc.geo.pdf.
36. Zakharov M. DataparkSearch at TREC 2005. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/datapark.geo.pdf.
37. Eppig JT, et al., The Mouse Genome Database (MGD): from genes to mice - a community resource for mouse biology. Nucleic Acids Research, 2005. 33: D471-D475.
38. Anonymous, The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research, 2004. 32: D258-D261.

39. Krupke D, et al., The Mouse Tumor Biology Database: integrated access to mouse cancer biology data. Experimental Lung Research, 2005. 31: 259-270.
40. Hill DP, et al., The mouse Gene Expression Database (GXD): updates and enhancements. Nucleic Acids Research, 2004. 32: D568-D571.
41. Strivens M and Eppig JT, Visualizing the laboratory mouse: capturing phenotype information. Genetica, 2004. 122: 89-97.
42. Si L and Kanungo T. Thresholding strategies for text classifiers: TREC 2005 biomedical triage task experiments. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/carnegie-mu-kanungo.geo.pdf.
43. Lam W, Han Y, and Chan K. Pattern-based customized learning for TREC Genomics Track categorization task. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/chineseu-hongkong-lam.geo.pdf.
44. Dayanik A, et al. DIMACS at the TREC 2005 Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/rutgersu-dimacs.geo.pdf.
45. Zheng ZH, et al. Applying probabilistic thematic clustering for classification in the TREC 2005 Genomics Track. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/queensu.geo.pdf.
46. Subramaniam LV, Mukherjea S, and Punjani D. Biomedical document triage: automatic classification exploiting category specific knowledge. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/ibm-india.subramaniam.geo.pdf.
47. Brow T, Settles B, and Craven M. Classifying biomedical articles by making localized decisions. The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). 2005. Gaithersburg, MD: National Institute for Standards & Technology. http://trec.nist.gov/pubs/trec14/papers/uwisconsin.geo.pdf.

HARD Track Overview in TREC 2005
High Accuracy Retrieval from Documents

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

1 Introduction

TREC 2005 saw the third year of the High Accuracy Retrieval from Documents (HARD) track. The HARD track explores methods for improving the accuracy of document retrieval systems, with particular attention paid to the start of the ranked list. Although it has done so in a few different ways in the past, budget realities limited the track to "clarification forms" this year. The question investigated was whether highly focused interaction with the searcher be used to improve the accuracy of a system. Participants created "clarification forms" generated in response to a query—and leveraging any information available in the corpus—that were filled out by the searcher. Typical clarification questions might ask whether some titles seem relevant, whether some words or names are on topic, or whether a short passage of text is related.

The following summarizes the changes from the HARD track in TREC 2004 [Allan, 2005]:

- There was no passage retrieval evaluation as part of the track this year.
- There was no use of metadata this year.
- The evaluation corpus was the full AQUAINT collection. In HARD 2003 the track used part of AQUAINT plus additional documents. In HARD 2004 it was a collection of news from 2003 collated especially for HARD.
- The topics were selected from existing TREC topics. The same topics were used by the Robust track [Voorhees, 2006]. The topics had not been judged against the AQUAINT collection, though had been judged against a different collection.
- There was no notion of "hard relevance" and "soft relevance", though documents were judged on a trinary scale of not relevant, relevant, or highly relevant.
- Clarification forms were allowed to be much more complex this year.
- Corpus and topic development, clarification form processing, and relevance assessments took place at NIST rather than at the Linguistic Data Consortium (LDC).
- The official evaluation measure of the track was R-precision.

The HARD track's Web page may also contain useful pointers, though is not guaranteed to be in place indefinitely. As of early 2006, it was available at http://ciir.cs.umass.edu/research/hard.

For TREC 2006, the HARD track is being "rolled into" the Question Answering track. The new aspect of the QA track is called "ciQA" for "complex, interactive Question Answering." The goal of ciQA is to investigate interactive approaches to cope with complex information needs specified by a templated query.

2 The Process

The HARD track proceeded as follows. This process follows roughly that of past years' tracks, though it simpler because passage retrieval was not an issue.

At the end of May, the track guidelines were finalized. Sites knew then that the evaluation corpus would be the AQUAINT collection (see Section 4), so could begin indexing the data and/or training their systems (see Section 7).

On June 15, 2005, participating sites received the set of 50 test topics from NIST (see Section 5).

Three weeks later, on July 7, sites had to submit the "baseline" ranked lists produced by their system (see Section 8). These runs ideally represented the best that the sites could do with only "classic" TREC topic information.

On the same day, sites were permitted to submit sets of clarification forms, where each set contained a form for each topic in the test set. The clarification form could contain almost anything that the site felt an answer would be useful for improving the accuracy of the query (e.g., possibly relevant passages, keywords that might reflect relevance). See Section 9 for more details.

For the next two weeks, assessors at NIST filled out clarification forms for the topics. On July 25, the clarification form responses were shipped to the sites.

On August 8, the sites submitted new "final" ranked lists that utilized information from the clarification forms (see Section 10).

Between then and early September, the assessors judged documents for relevance (see Section 6). Relevance assessments ("qrels") were made available to the researchers on September 9, 2005.

3 Participation

A total of 16 sites submitted 122 runs for the track. The following breakdown shows how many runs each site submitted, broken down by baseline and final runs, as well as the number of clarification forms submitted.

# runs			
Base	Final	# CFs	Participating site
0	10	2	Chinese Academy of Sciences
1	8	2	Chinese Academy of Sciences NLPR
4	6	2	Indiana University
2	7	2	Meiji University
1	11	2	Rutgers University
2	6	2	SAIC/U. of Virginia
1	1	1	University College Dublin
1	6	3	University of Illinois at Urbana-Champaign
3	3	1	University of Maryland, College Park
4	4	2	University of Massachusetts
1	3	3	University of North Carolina
2	4	2	University of Pittsburgh
1	7	2	University of Strathclyde
2	6	2	University of Twente
2	4	3	University of Waterloo
3	5	3	York University

4 HARD Corpus

For TREC 2005, the HARD track used the AQUAINT corpus. That corpus is available from the Linguistic Data Consortium for a modest fee, and was made available to HARD participants who were not a member of the LDC for no charge. The LDC's description of the corpus[1] is:

> The AQUAINT Corpus, Linguistic Data Consortium (LDC) catalog number LDC2002T31 and isbn1-58563-240-6 consists of newswire text data in English, drawn from three sources: the Xinhua News Service (People's Republic of China), the New York Times News Service, and the Associated Press Worldstream News Service. It was prepared by the LDC for the AQUAINT Project, and will be used in official benchmark evaluations conducted by National Institute of Standards and Technology (NIST).

The corpus is roughly 3Gb of text and includes 1,033,461 documents (about 375 million words of text, according to the LDC's web page). All documents in the collection were used for the HARD evaluation.

5 Topics

Topics were selected from among existing TREC topics that almost no system was able to handle well in previous years. Because those old topics were to be judged on a new corpus (AQUAINT), they were manually vetted to ensure that at least three relevant documents existed in the AQUAINT corpus. These topics were also used by the TREC 2005 Robust track [Voorhees, 2006].

The topic numbers used were: 303, 307, 310, 314, 322, 325, 330, 336, 341, 344, 345, 347, 353, 354, 362, 363, 367, 372, 374, 375, 378, 383, 389, 393, 394, 397, 399, 401, 404, 408, 409, 416, 419, 426, 427, 433, 435, 436, 439, 443, 448, 622, 625, 638, 639, 648, 650, 651, 658, and 689.

6 Relevance judgments

Topics were judged for relevance by the same assessor who answered the clarification forms for the topic (see Section 9 for more information on clarification forms). In the first two years of HARD, that same person also created the original topic statement; however, because topics were re-used, it was not possible to use the same person for the original step. No attempt was made to ensure that this year's assessor's notion of relevance would match that of the original assessor.

Six assessors worked on the fifty topics, as follows:

Assessor A: 347 399 401 404 408 409 419 426
Assessor B: 625 638 639 648 650 651 658 689
Assessor C: 427 433 435 436 439 443 448 622
Assessor D: 303 322 345 354 362 363 367 383 393
Assessor E: 336 341 353 372 375 378 394 397
Assessor F: 307 310 314 325 330 344 374 389 416

Documents were judged as either not relevant, relevant, or highly relevant. For purposes of this track, judgments of *relevant* and *highly relevant* were treated as the same.

[1] At http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T31 as of May 2006.

7 Training data

The data collections from the HARD tracks of TREC 2003 [Allan, 2004] and 2004 [Allan, 2005] were available for training. All of that data was made available to HARD track participants courtesy of the Linguistic Data Consortium. The data was provided for use only in the HARD 2005 evaluation with the expectation that it will be destroyed at the completion of the track (i.e., after the final papers are written). The HARD 2004 corpus and topics are now available for purchase from the LDC as catalogue numbers LDC2005T28 and LDC2005T29[2].

The TREC 2004 HARD track used a corpus of news from 2003, had 49 topics with several metadata fields. Topics, relevance judgments, and clarification forms were provided.

The TREC 2003 HARD track corpus was a set of 372,219 documents totaling 1.7Gb from the 1999 portion of the AQUAINT corpus, along with some US government documents from the same year (Congressional Record and Federal Register). The topics were somewhat like standard TREC topics, but included lots of searcher and query metadata. Topics, relevance judgments, and clarification forms were provided.

8 Baseline submissions

Submissions of baseline runs were in the standard TREC submission format used for ad-hoc queries. Up to 1000 documents were provided in rank order for each of the 50 topics. The details were in a file with lines containing a topic number, a document ID, the document's rank against that topic, and its score (along with some other bits of bookkeeping information). Every topic was required to have at least one document retrieved, and it could have anywhere from one to 1,000 documents.

Sites were asked to provide the following information:

1. *Was this an entirely automatic run or a manual run?* Two baseline runs were manual, all others were automatic.
2. *Did you use the title, description, and/or narrative fields for this run?* The runs included 9 using just the title field, 3 using just description, 8 combining title and description, and 10 also adding in the narrative.
3. *To what extent did you use earlier relevance judgments on the topics?* One run claimed to have used the judgments of these topics against prior TREC corpora.
4. *A short description of the run.*
5. *Preference in terms of judging of this run?* Only one baseline run per site was included in the judging pool.

9 Clarification forms

All 16 participating sites submitted at least one clarification forms: two submitted one form, ten submitted two forms, and four sites submitted three. All submitted forms were filled out, even though the track guidelines only guaranteed that two would be.

Clarification forms were filled out by the NIST assessors using the following platform:

[2]Described at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T28 and ... LDC2005T29, respectively, as of May 2006.

- Redhat Enterprise Linux, version "3 workstation"
- 20-inch LCD monitor with 1600x1200 resolution, true color (millions of colors)
- Firefox Web browser, v1.0.3
- No assumption that the machine is connected to any network at all. (The goal was to have it disconnected from all networks of any sort, but that proved infeasible in the NIST environment.)

In past years, the contents of the clarification forms were strictly controlled to allow only a limited subset of HTML. This year, virtually all restrictions were lifted, meaning that sites could include Javacript, Java, images, or the like. The following restrictions were made:

- The forms had to assume they were running on a computer that is disconnected from all networks, so all necessary information had to be included as part of the form. If it required multiple files, they all had to be within the same directory structure. Sites could not assume that all of its clarification forms would be on the same computer.
- It was not possible to invoke any cgi-bin scripts
- It was not possible to write to disk

Clarification forms could be presented in almost any layout, but had to include the following items:

- **<form action="/cgi-bin/clarification_submit.pl" method="post">**
  This indicates the script where the output was generated (all it did was output the selected information).
- **<input type="hidden" name="site" value="XXXXn">**
  Here, "XXXX" is a 4-letter code designating the site (provided in the lead-up to the baseline submission) and "n" is a run number. The run numbers reflected the priority order of the form. That is, XXXX1 will be processed then XXXX2 and so on.
- **<input type="hidden" name="topicid" value="000">**
  Indicates the topic number, a 3-digit code with zeros padding as needed (001 rather than 01 or 1).
- **<input type="submit" name="send" value="submit">**
  This is the submit button that had to appear somewhere on the page.

In addition, sites were strongly encouraged to include somewhere on the page the topic number (e.g., "303") and the title of the topic to provide a sanity check that the annotators were, indeed, answering the correct questions.

For each submission, all clarification forms were put in a single directory (folder) with the name indicated (e.g., NIST1). Each clarification form inside that directory was also a directory with the name of the submission and the topic number (e.g., NIST1_043 for topic 43 of the NIST1 submission).

Inside *that* directory, the main clarification form was called index.html. It could access any files from within the directory hierarchy, using relative pathnames. For example, "logo.gif" would refer to the file NIST1/NIST1_043/logo.gif within the directory structure, and "../logo.gif" would refer to NIST1/logo.gif".

Sites were asked the following information about each submitted form:

1. Did you use clustering to generate this form?
2. Did you use text summarization, either extractive or generative?

3. Did you use document-level feedback? That is, did you ask the user to judge an entire document for relevance, even if you did so using a title, passage, or keywords from the document?

4. Did you ask the user to judge selected passages of text, independent of the documents they came from?

5. Did you ask the user to judge keywords for relevance, independent of the documents they came from?

6. If you used any techniques not listed above, briefly list them at the bullet-list level of detail.

7. Did you use any sources of information beyond the query and AQUAINT corpus and, if so, what were they?

The assessors spent no more than three minutes per form no matter how complex the form was. The three minutes included time needed to load the form, initialize it, and do any rendering, so unusually complex or large forms were implicitly penalized. At the end of three minutes, if the assessor had not pressed the "submit" button, the form was timed out and forcibly submitted (anything entered up to that point was saved).

NIST recorded the time spent on the form returned for each form. That information was returned in a separate file along with all of the clarification form responses. Assessors were never permitted more than 180 seconds per form, but some of the reported times were greater than 180 because of the time it took for the system to "shut down" a form if the time limit expired.

Clarification forms were presented to annotators in an order to minimize the chance that one form would adversely (or positively) impact the use of another form. Tables 1 and 2 shows the rotation that was used for the submitted clarification forms.

10 Final submissions

Final submissions incorporated information gleaned from clarification forms and combined that with any other retrieval techniques to achieve the best run possible.

The following questions were asked for each submission:

1. *Which of your baseline runs is an appropriate baseline?* There were 26 submissions that indicated that the final run did not have a corresponding baseline run. This often reflected a site's providing a new "baseline" or trying out a technique that was developed after the baseline runs and so had no corresponding baseline.

2. *Which of your clarification forms was used to generated this final run?* There were 33 final runs that indicated they did *not* use a clarification form.

3. *Other than the clarification form's being answered, was this an entirely automatic run or a manual run?* Only four of the final runs were marked as being manual runs; the remaining 88 were automatic.

4. *Did you use the title, description, and/or narrative fields for this run?* Here, 28 runs used just the title, 2 used just the description, 39 combined the title and description, and 23 also included the narrative.

5. *To what extent did you use earlier relevance judgments on the topics?* A total of 13 runs indicated that they used the earlier relevance judgments.

6. *A short description of the run.*

7. *What is the preference in terms of judging of this run?* Only one final run from each site was included in the judging pool.

	NCAR1	MARY1	INDI2	STRA2	UIUC3	UIUC1	NCAR3	TWEN2	PITT1	YORK2	CASP1	CASS2	NCAR2	PITT2	MASS1	SAIC1	YORK1
T1	28	30	23	5	19	3	20	12	15	16	2	17	32	22	29	7	21
T2	29	31	24	6	20	4	21	13	16	17	3	18	33	23	30	8	22
T3	30	32	25	7	21	5	22	14	17	18	4	19	34	24	31	9	23
T4	31	33	26	8	22	6	23	15	18	19	5	20	1	25	32	10	24
T5	32	34	27	9	23	7	24	16	19	20	6	21	2	26	33	11	25
T6	33	1	28	10	24	8	25	17	20	21	7	22	3	27	34	12	26
T7	34	2	29	11	25	9	26	18	21	22	8	23	4	28	1	13	27
T8	1	3	30	12	26	10	27	19	22	23	9	24	5	29	2	14	28
T9	2	4	31	13	27	11	28	20	23	24	10	25	6	30	3	15	29
T10	3	5	32	14	28	12	29	21	24	25	11	26	7	31	4	16	30
T11	4	6	33	15	29	13	30	22	25	26	12	27	8	32	5	17	31
T12	5	7	34	16	30	14	31	23	26	27	13	28	9	33	6	18	32
T13	6	8	1	17	31	15	32	24	27	28	14	29	10	34	7	19	33
T14	7	9	2	18	32	16	33	25	28	29	15	30	11	1	8	20	34
T15	8	10	3	19	33	17	34	26	29	30	16	31	12	2	9	21	1
T16	9	11	4	20	34	18	1	27	30	31	17	32	13	3	10	22	2
T17	10	12	5	21	1	19	2	28	31	32	18	33	14	4	11	23	3
T18	11	13	6	22	2	20	3	29	32	33	19	34	15	5	12	24	4
T19	12	14	7	23	3	21	4	30	33	34	20	1	16	6	13	25	5
T20	13	15	8	24	4	22	5	31	34	1	21	2	17	7	14	26	6
T21	14	16	9	25	5	23	6	32	1	2	22	3	18	8	15	27	7
T22	15	17	10	26	6	24	7	33	2	3	23	4	19	9	16	28	8
T23	16	18	11	27	7	25	8	34	3	4	24	5	20	10	17	29	9
T24	17	19	12	28	8	26	9	1	4	5	25	6	21	11	18	30	10
T25	18	20	13	29	9	27	10	2	5	6	26	7	22	12	19	31	11
T26	19	21	14	30	10	28	11	3	6	7	27	8	23	13	20	32	12
T27	20	22	15	31	11	29	12	4	7	8	28	9	24	14	21	33	13
T28	21	23	16	32	12	30	13	5	8	9	29	10	25	15	22	34	14
T29	22	24	17	33	13	31	14	6	9	10	30	11	26	16	23	1	15
T30	23	25	18	34	14	32	15	7	10	11	31	12	27	17	24	2	16
T31	24	26	19	1	15	33	16	8	11	12	32	13	28	18	25	3	17
T32	25	27	20	2	16	34	17	9	12	13	33	14	29	19	26	4	18
T33	26	28	21	3	17	1	18	10	13	14	34	15	30	20	27	5	19
T34	27	29	22	4	18	2	19	11	14	15	1	16	31	21	28	6	20
T35	28	30	23	5	19	3	20	12	15	16	2	17	32	22	29	7	21
T36	29	31	24	6	20	4	21	13	16	17	3	18	33	23	30	8	22
T37	30	32	25	7	21	5	22	14	17	18	4	19	34	24	31	9	23
T38	31	33	26	8	22	6	23	15	18	19	5	20	1	25	32	10	24
T39	32	34	27	9	23	7	24	16	19	20	6	21	2	26	33	11	25
T40	33	1	28	10	24	8	25	17	20	21	7	22	3	27	34	12	26
T41	34	2	29	11	25	9	26	18	21	22	8	23	4	28	1	13	27
T42	1	3	30	12	26	10	27	19	22	23	9	24	5	29	2	14	28
T43	2	4	31	13	27	11	28	20	23	24	10	25	6	30	3	15	29
T44	3	5	32	14	28	12	29	21	24	25	11	26	7	31	4	16	30
T45	4	6	33	15	29	13	30	22	25	26	12	27	8	32	5	17	31
T46	5	7	34	16	30	14	31	23	26	27	13	28	9	33	6	18	32
T47	6	8	1	17	31	15	32	24	27	28	14	29	10	34	7	19	33
T48	7	9	2	18	32	16	33	25	28	29	15	30	11	1	8	20	34
T49	8	10	3	19	33	17	34	26	29	30	16	31	12	2	9	21	1
T50	9	11	4	20	34	18	1	27	30	31	17	32	13	3	10	22	2

Table 1: Rotation used to fill out clarification forms (the right edge of the table continues in Table 2). The rows of the table correspond to topics and the columns to clarification forms from sites. For example, the form indicates that NCAR's primary clarification form (NCAR1) will be the 28th considered for topic 1, the 29th for topic 2, ..., the 1st for topic 8, and so on. Similarly, for topic 1, the assessor first did INDI1's form (see Table 1), then that for CASP1, then UIUC1's, followed by MEIJ1's, and so on.

11 Overview of submissions

As mentioned above, 16 sites participated. The following statistics provide some details of the submissions. Note that the information is largely self-reported and has not been rigorously verified, so it is possible that it may be somewhat inaccurate.

A total of 30 baseline runs were submitted from 15 sites. One of those 15 sites made use of the earlier judgments for the topics (on a different corpus and using a different assessor).

A total of 35 sets of clarification forms were submitted. The average time per form on a single question was 116.5 seconds, with a minimum of five seconds and a maximum of 180 seconds. (In fact, one query's form reported taking 676 seconds, but the more than 8 additional minutes were presumably consumed by the system trying to force the form to close after the three minutes had expired.)

Every site had at least one form that took the full three minutes, and many had a dozen or two that took that long. The University of Massachusetts had the distinction of being the only site that used the full

	CASS1	DU8L1	UWAT1	MASS2	CASP2	STRA3	UWAT2	MEIJ1	MEIJ2	RUTG2	YORK3	RUTG1	SAIC2	INDI1	TWEN1	UIUC2	UWAT3
T1	26	11	25	13	6	27	14	4	33	18	31	24	8	1	9	10	34
T2	27	12	26	14	7	28	15	5	34	19	32	25	9	2	10	11	1
T3	28	13	27	15	8	29	16	6	1	20	33	26	10	3	11	12	2
T4	29	14	28	16	9	30	17	7	2	21	34	27	11	4	12	13	3
T5	30	15	29	17	10	31	18	8	3	22	1	28	12	5	13	14	4
T6	31	16	30	18	11	32	19	9	4	23	2	29	13	6	14	15	5
T7	32	17	31	19	12	33	20	10	5	24	3	30	14	7	15	16	6
T8	33	18	32	20	13	34	21	11	6	25	4	31	15	8	16	17	7
T9	34	19	33	21	14	1	22	12	7	26	5	32	16	9	17	18	8
T10	1	20	34	22	15	2	23	13	8	27	6	33	17	10	18	19	9
T11	2	21	1	23	16	3	24	14	9	28	7	34	18	11	19	20	10
T12	3	22	2	24	17	4	25	15	10	29	8	1	19	12	20	21	11
T13	4	23	3	25	18	5	26	16	11	30	9	2	20	13	21	22	12
T14	5	24	4	26	19	6	27	17	12	31	10	3	21	14	22	23	13
T15	6	25	5	27	20	7	28	18	13	32	11	4	22	15	23	24	14
T16	7	26	6	28	21	8	29	19	14	33	12	5	23	16	24	25	15
T17	8	27	7	29	22	9	30	20	15	34	13	6	24	17	25	26	16
T18	9	28	8	30	23	10	31	21	16	1	14	7	25	18	26	27	17
T19	10	29	9	31	24	11	32	22	17	2	15	8	26	19	27	28	18
T20	11	30	10	32	25	12	33	23	18	3	16	9	27	20	28	29	19
T21	12	31	11	33	26	13	34	24	19	4	17	10	28	21	29	30	20
T22	13	32	12	34	27	14	1	25	20	5	18	11	29	22	30	31	21
T23	14	33	13	1	28	15	2	26	21	6	19	12	30	23	31	32	22
T24	15	34	14	2	29	16	3	27	22	7	20	13	31	24	32	33	23
T25	16	1	15	3	30	17	4	28	23	8	21	14	32	25	33	34	24
T26	17	2	16	4	31	18	5	29	24	9	22	15	33	26	34	1	25
T27	18	3	17	5	32	19	6	30	25	10	23	16	34	27	1	2	26
T28	19	4	18	6	33	20	7	31	26	11	24	17	1	28	2	3	27
T29	20	5	19	7	34	21	8	32	27	12	25	18	2	29	3	4	28
T30	21	6	20	8	1	22	9	33	28	13	26	19	3	30	4	5	29
T31	22	7	21	9	2	23	10	34	29	14	27	20	4	31	5	6	30
T32	23	8	22	10	3	24	11	1	30	15	28	21	5	32	6	7	31
T33	24	9	23	11	4	25	12	2	31	16	29	22	6	33	7	8	32
T34	25	10	24	12	5	26	13	3	32	17	30	23	7	34	8	9	33
T35	26	11	25	13	6	27	14	4	33	18	31	24	8	1	9	10	34
T36	27	12	26	14	7	28	15	5	34	19	32	25	9	2	10	11	1
T37	28	13	27	15	8	29	16	6	1	20	33	26	10	3	11	12	2
T38	29	14	28	16	9	30	17	7	2	21	34	27	11	4	12	13	3
T39	30	15	29	17	10	31	18	8	3	22	1	28	12	5	13	14	4
T40	31	16	30	18	11	32	19	9	4	23	2	29	13	6	14	15	5
T41	32	17	31	19	12	33	20	10	5	24	3	30	14	7	15	16	6
T42	33	18	32	20	13	34	21	11	6	25	4	31	15	8	16	17	7
T43	34	19	33	21	14	1	22	12	7	26	5	32	16	9	17	18	8
T44	1	20	34	22	15	2	23	13	8	27	6	33	17	10	18	19	9
T45	2	21	1	23	16	3	24	14	9	28	7	34	18	11	19	20	10
T46	3	22	2	24	17	4	25	15	10	29	8	1	19	12	20	21	11
T47	4	23	3	25	18	5	26	16	11	30	9	2	20	13	21	22	12
T48	5	24	4	26	19	6	27	17	12	31	10	3	21	14	22	23	13
T49	6	25	5	27	20	7	28	18	13	32	11	4	22	15	23	24	14
T50	7	26	6	28	21	8	29	19	14	33	12	5	23	16	24	25	15

Table 2: Continuation of Table 1; this table appears to the right of that table.

three minutes of annotator time for *every* form. Those forms were apparently designed to collect as much information as possible during clarification time for later processing to determine which questions were most useful [Diaz and Allan, 2006].

A total of 92 final runs were submitted across the 16 sites. Of those, three runs made use of the past judgments. Different sites used different parts of the topics for their runs:

- 28 runs were title-only queries
- 2 runs were description-only queries
- 38 runs combined the title and description
- 24 runs included the narrative along with the title and description

All runs were automatic (not counting clarification form interaction) except for those submitted by the University of Maryland, where experiments used a trained intermediary to collect potentially useful information for a clarification form [Lin et al., 2006].

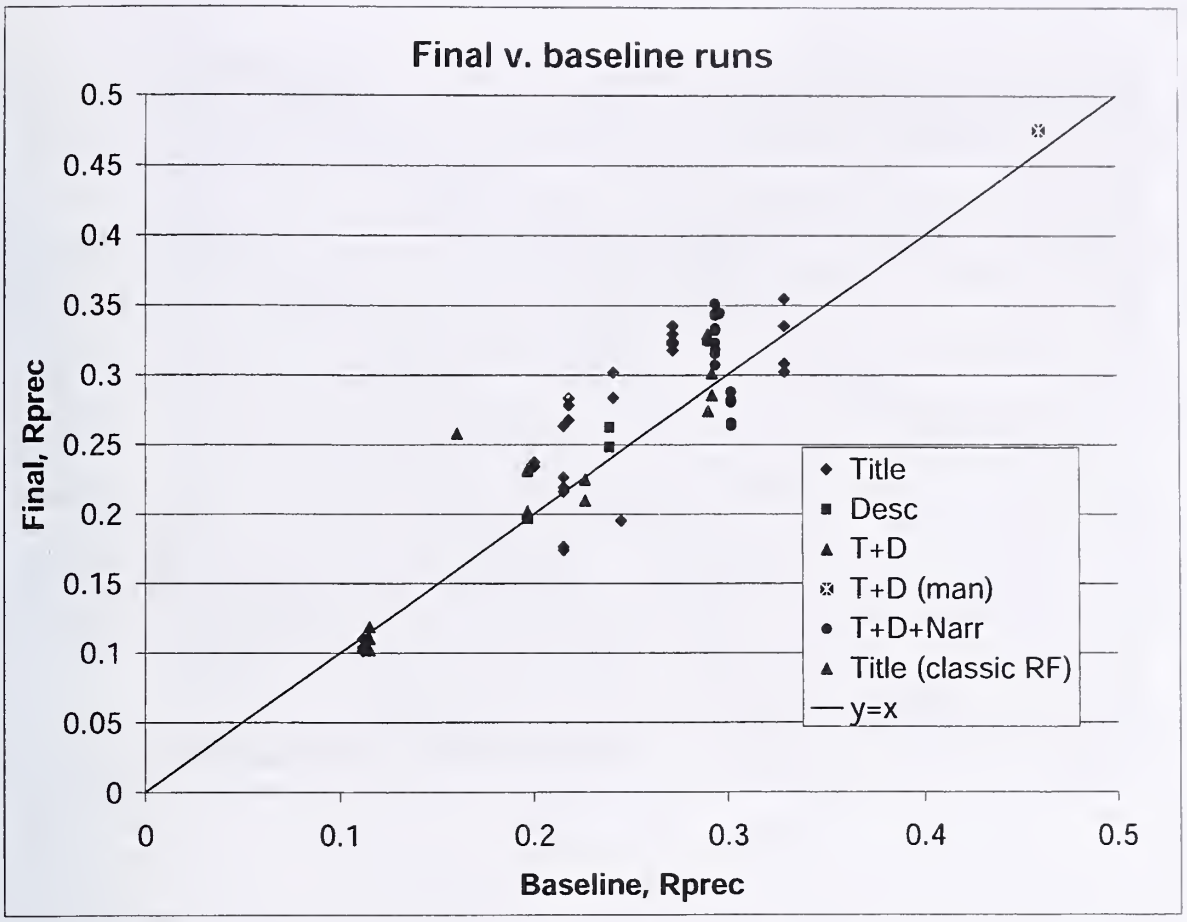

Figure 1: Comparison of R-precision values in baseline runs and runs after using a clarification form (only runs that identified a corresponding baseline run are included).

12 Discussion

System output was evaluated by R-precision, defined as precision at R documents retrieved, where R is the number of known relevant documents in the collection.

Figure 1 shows overall performance as impacted by clarification forms. Recall that when a final run was submitted, sites were asked to indicate which of their baseline runs was used as a starting point. The graph includes a point for each such baseline-final pair. Because (by chance) different baseline runs never had the same score, points that make up vertical lines represent multiple final runs that used the same baseline run. For example, the run at baseline R-precision of 0.3291 was used for four final runs that had R-precision ranging from 0.3024 to 0.3547.

Point colors and shape reflect which portions of the topic were used for the query, though the differences may not be easily visible in a grayscale print. The (excellent) outlier labeled "T+D (man)" in the upper right is the manual run from the University of Maryland. The red triangle at baseline 0.1599 and final 0.2581, labeled "Title (classic RF)", is a special run created by NIST, and is discussed further below.

12.1 General observations

Just considering baseline runs, the automatic runs had R-precision scores ranging from 0.1116 to 0.3291. Using the title, description, and narrative seemed to be helpful, since four different sites achieved comparable scores. However, some baselines without the narrative performed just as well, and the best automatic baseline used only the title.

Ultimately, the goal of the HARD track was to explore the value added by clarification forms. That means that it is the improvement from baseline to final that is more interesting. In the graph, points below the $y = x$ line had final runs that were worse than their corresponding baseline runs; those above the line improved. Most of the sites were able to improve on their baseline performance.

12.2 Classic relevance feedback

In previous years of the HARD track, there was a concern that simple relevance feedback of documents might be a simpler and more effective type of clarification form. To explore that issue this year, NIST volunteered to provide a form that was purely relevance feedback. To do that, NIST ran a baseline system and then created a clarification form that included the top-ranked documents, asking that they be judged as relevant or not. The baseline system was Prise3, a system based on the Lucene open source IR engine, so it used a tf-idf style of retrieval. Prise3 was the same system used to create new topics for other tracks this year. The title field to retrieve the top 50 documents.

The clarification form listed, along with the query's title and description, the title of the top 50 documents. The assessor could click on a link to see the full text of a document if needed. The assessor used his or her three minutes to judge as many documents as possible, and then a new query was created using that information. Because Prise3 did not support relevance feedback at that time, the final run version 11.0 of the well-known SMART system. The system was tuned using the Robust 2004 topics on the past corpus, without paying any special attention ot the topics that were re-used for HARD this year. The tuning used the top-ranked five relevant documents on that corpus, as an estimate of what might come back from the clarification forms. The tuned parameters were the weighting scheme (ltc.lnc), the number of feedback terms to select (50), and the Rocchio parameters ($\alpha = 4, \beta = 2$).

The red triangle in Figure 1 shows the performance of the NIST feedback runs, where the baseline performance is the Prise3 system and the final performance is for the SMART system. The point is dramatically above the $y = x$ line, showing the dramatic improvement (more than 60%) this approach can cause. Unfortunately, the baseline was substantially below the better-performing systems, making it difficult to know whether simple relevance feedback would be equally effective at different qualities of baseline system. The results suggest that having a "pure" relevance feedback clarification form from every system might be a useful point for comparison.

12.3 Results by query

To a limited degree, it appears that better performing baselines result in larger gains from the clarification form. Figure 2 shows a breakdown of the same runs with each query represented. The graph shows a clear suggestion that it is easier to improve better-performing queries, but also demonstrates that poor-performing queries can be improved and have more *room* for improvement.

Another way of looking at the same question is to explore the absolute gain as a function of baseline R-precision. Figure 3 shows the same queries as Figure 2, but the y-axis shows the gain rather than value of R-precision. There is a very slight trend toward lower gain given higher baseline R-precision, but the fit is poor and the slope is almost horizontal. The graph suggests a strong negative correlation to the eye, but it

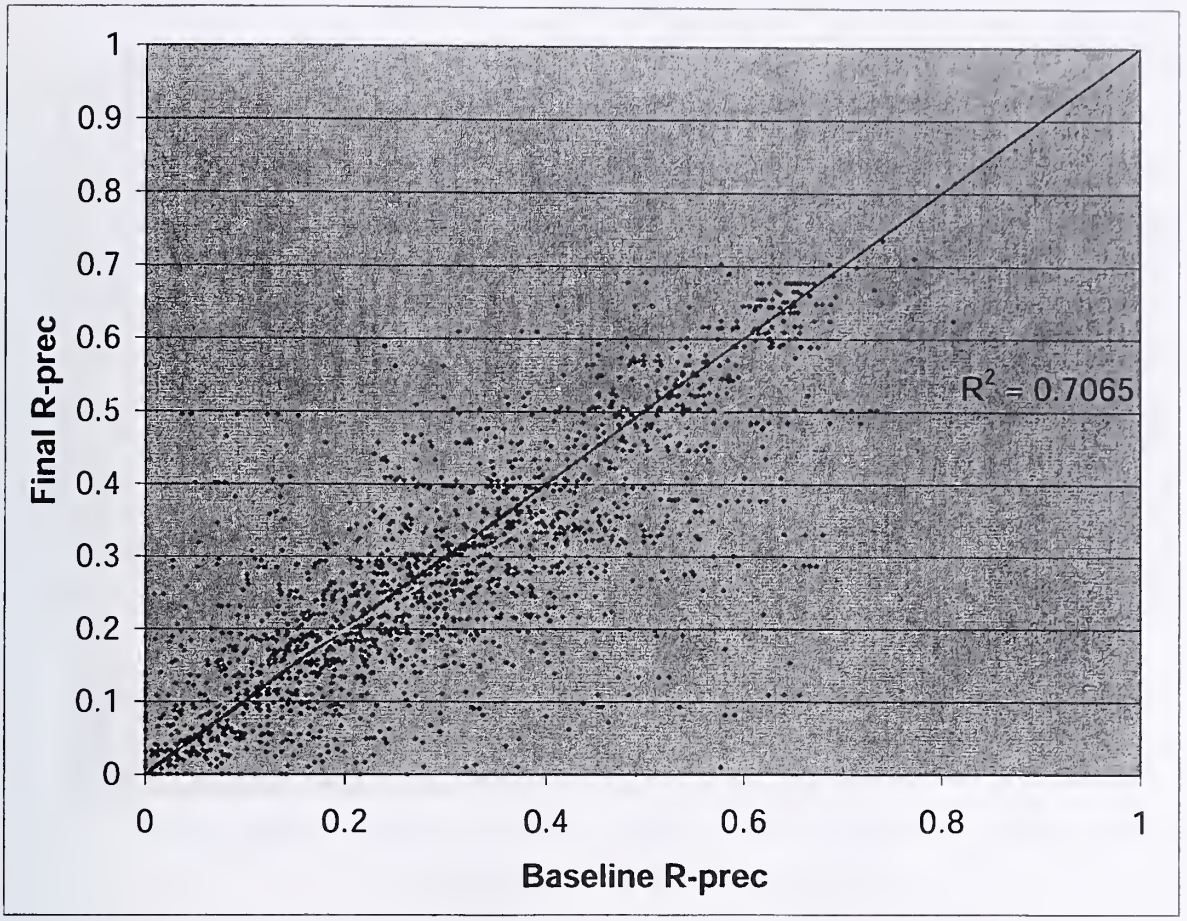

Figure 2: Comparison of R-precision values in baseline and final runs on a query-by-query basis. Each query from each run pair represented in Figure 1 is represented by a point. The $y = x$ line is shown as well as a linearly fit trend line.

is an artifact of the absolute loss being capped by the value of baseline R-precision—that is, if the baseline R-precision is 0.02, it is not possible to lose more than 0.02, but the gain can be quite large.

Figure 4 shows the absolute gain as a function of the number of relevant documents in the query. Again, there is a very weak trend toward more gain given more relevant documents in the pool. But the graph very clearly shows that the variance of the gain is large across all queries, regardless of the number of relevant documents they have.

Finally we consider the possibility that gain is correlated with the amount of time spent in clarification forms. Figure 5 shows that having annotators spend more time providing clarification information did not in and of itself increase realized gain. (Any effect may be obscured because a third of the interactions with annotators were truncated at 180 seconds, meaning we do not know how much time they actually might have spent.)

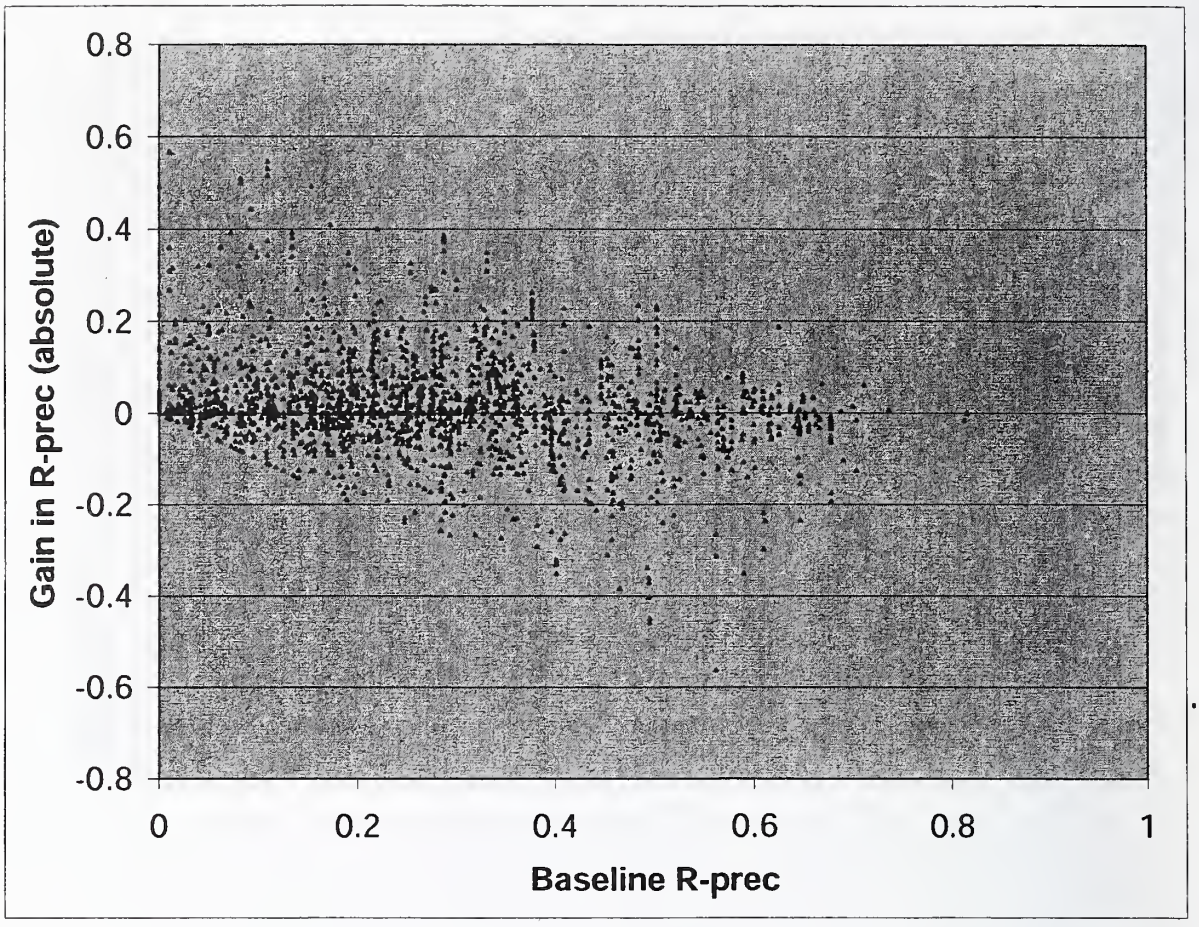

Figure 3: Comparing absolute gain in R-precision to baseline R-precision value.

12.4 Comparing two individual runs

It is illuminating to compare two runs that did well in the overall evaluation. We will consider the top performing title-only queries from two different groups.

1. Run MASStrmS is the automatic run with highest final R-precision. It started with baseline run MASSbaseTEE3 that had R-precision of 0.3291. It incorporated information from clarification form MASS1 and then achieved a final R-precision of 0.3547. That represents a 0.0256 gain in R-precision, an 8% relative improvement.

2. Run UIUChCFB3 is the automatic run with second highest final R-precision. It started with baseline run UIUC05Hardb0 that had R-precision of 0.2723. It incorporated information from clarification form UIUC3 and then achieved a final R-precision of 0.3355. That represents a 0.0623 gain in R-precision, a 23% relative improvement.

Figure 6 shows scatterplots of baseline and final R-precision values for the two runs, with UMass' run on the left and UIUC's on the right. For most queries in the UMass results, the final runs are almost identical to the baseline runs. However, a handful of queries with very low baseline scores show remarkable improvement, accounting for most of the gain in that system. This run appears to represent a very conservative query modification strategy, a reasonable choice given the high quality baseline.

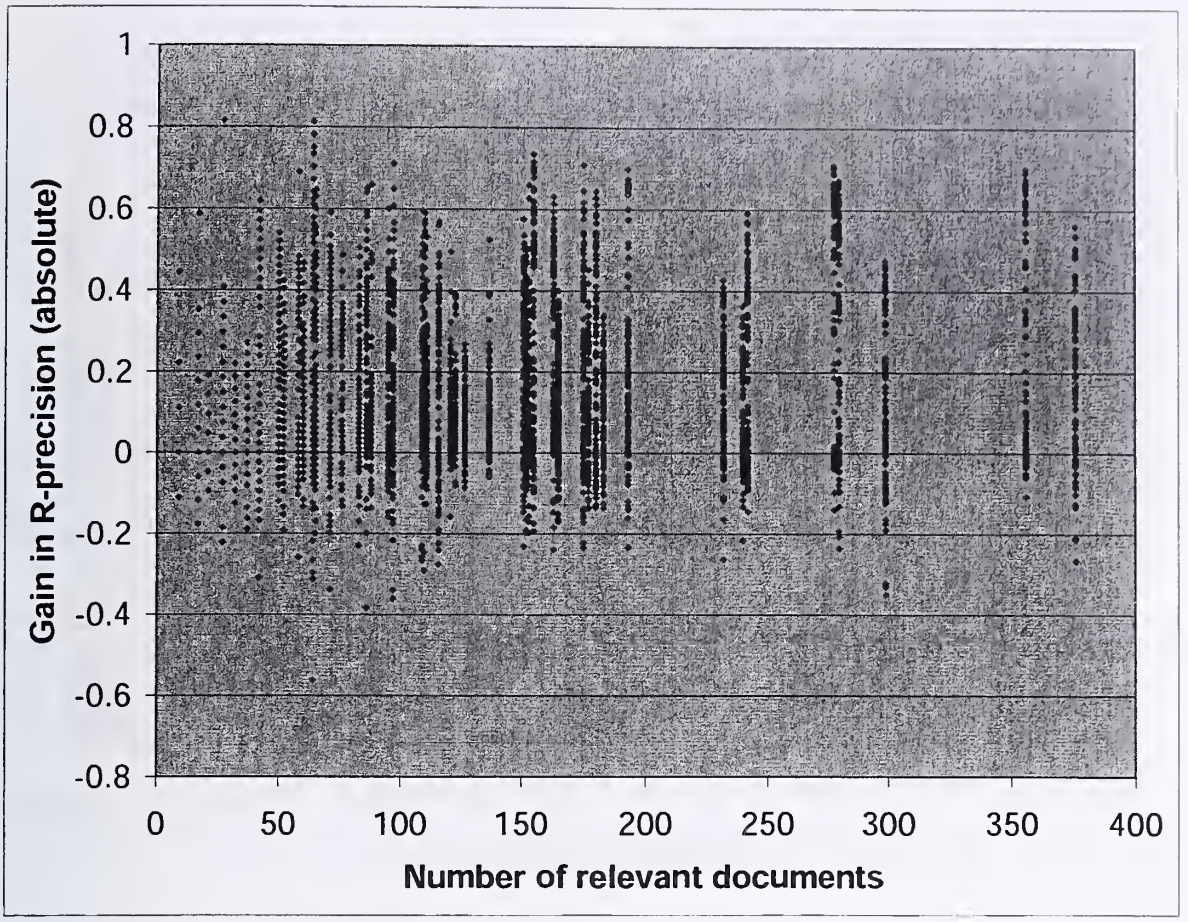

Figure 4: Comparing absolute gain in R-precision to the number of relevant documents for a query.

The UIUC run, in contrast, shows dramatic changes between the baseline and final runs. A large number of queries improve and a handful are significantly harmed. The strategy here is clearly much riskier and often pays off handsomely, trimming much of the baseline performance difference between UMass and UIUC.

Finally we do a direct comparison of how queries performed in the two systems. Figure 7 has an entry on the x-axis for every query. The queries are sorted by the final R-precision value of the query in the UIUChCFB3 run, the solid (blue) line that degrades smoothly from the upper left to the lower right. The corresponding baseline performance for that query is represented by (blue) diamonds.

The UMASStrmS final R-precision values are represented by the jagged (brown) line that roughly follows the trend of the UIUChCFB3 line, with the (red) triangles indicating baseline effectiveness.

Query effectiveness at the two sites follows a similar trend, but huge differences are common, with each site out-performing another by large margins in some cases. For example, query 651 show comparable baseline performance for the two sites, but successful clarification only by UIUC. Query 409 shows roughly the opposite result. Query 389 shows a case where UMass had substantially higher baseline performance, but UIUC's final run topped the UMass effectiveness by a good bit.

Comparing two systems provides only a glimpse of what is happening during clarification and final runs. It does suggest that different approaches work better for different queries, leading to the obvious question of whether it is possible to combine the clarification forms or to predict when one style is likely to be more

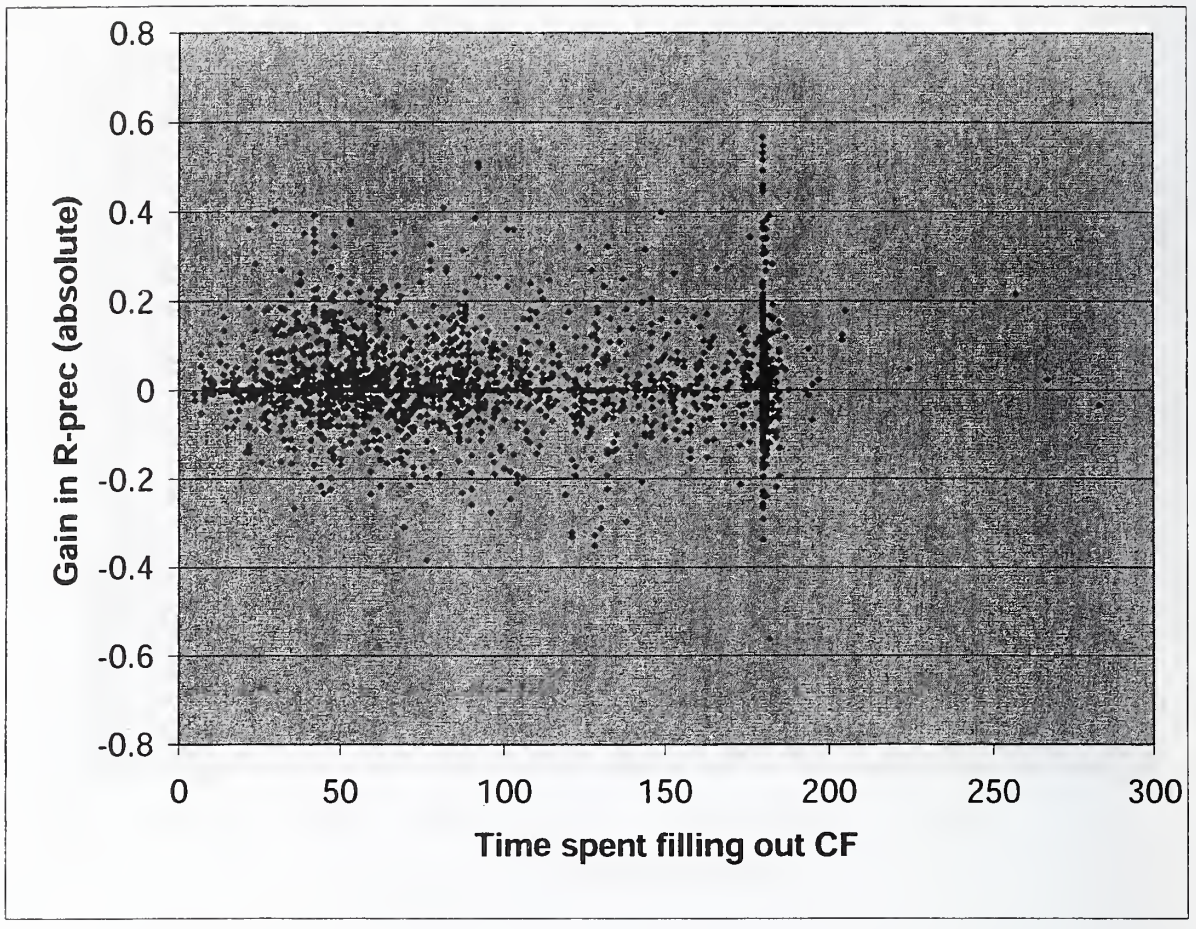

Figure 5: Comparing absolute gain in R-precision to time spent in the clarification form. Note that the density of scores at 180 seconds corresponds to the maximum time allowed in a form. The handful of scores beyond 180 seconds represent clarification forms that were difficult to "shut down" (see Section 9).

useful.

13 Conclusion

Several sites were able to show appreciable average gains from using clarification forms. None of the gains was consistently dramatic, however, begging the question of whether the time spent clarifying a query was a worthwhile investment. Further amplifying that question, it is worth pointing out that the best best automatic Robust track run beat all of the automatic baseline and final HARD track runs. (Of course, it is unknown whether a clarification form based on that run would improve the results further.)

This year there was an interesting variety of clarification forms tried. Forms of user-assisted query expansion were very popular, but sites also considered relationships between terms [Diaz and Allan, 2006, Yang et al., 2006], passage feedback [Diaz and Allan, 2006], incorporated summarization [Jin et al., 2006], and even used elaborate visualizations based on self organizing maps [He and Ahn, 2006]. The track itself did not provide clear support for any of these approaches.

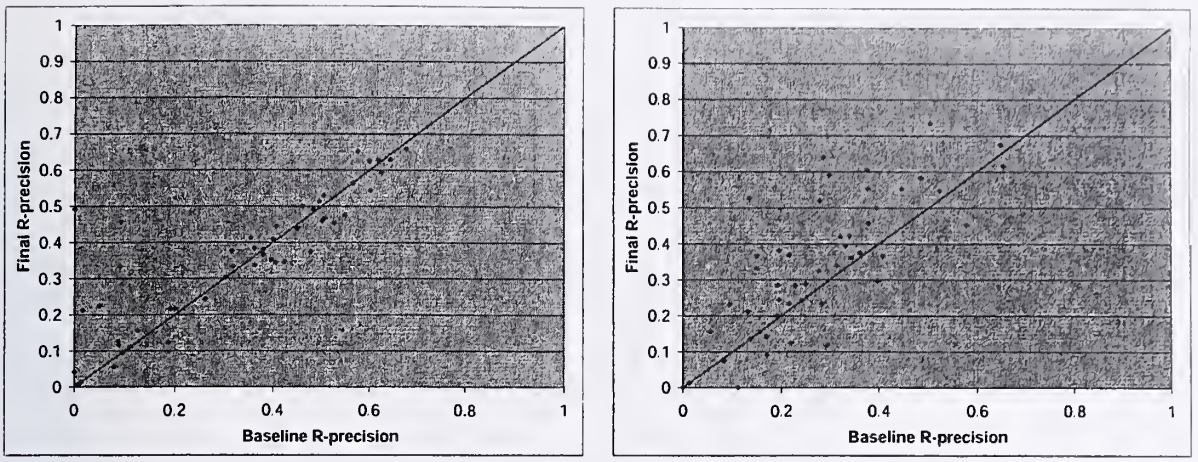

Figure 6: Comparison of baseline and final R-precision values for the MASStrmS run (left) and for the UIUChCFB3 run (right), broken down by query.

It is important to note that the clarification forms do not represent "interactive information retrieval" experiments. They provide a highly focused and very limited type of interaction that can (potentially) improve the effectiveness of document retrieval. Whether these clarification forms can be deployed in a way that pleases a user or that will actually be used is an entirely different question, one that would have to be tested in a more realistic environment.

After a three-year run, TREC 2005 was the end of the HARD track. For TREC 2006, it is being made part of the Question Answering track as "ciQA", or "complex, interactive Question Answering." The goal of ciQA is to investigate interactive approaches to cope with complex information needs specified by a templated query.

Acknowledgments

As do most tracks in TREC, the HARD track owes a debt of gratitude to Ellen Voorhees at NIST, who not only provided the typical TREC infrastructure, but who was an important contributor in the shaping of the track's goals and evaluation approaches. The Linguistic Data Consortium kindly provided training data to participants and words of wisdom to NIST organizers. Stephanie Strassel and Meghan Glenn of the LDC were particularly generous with their time and support.

Finally, the track would not have been successful without the involvement of its participants—not only by participating, but by designing the track during discussions in person and on-line. Diane Kelly at the University of North Carolina should be singled out for rapid production of the rotation tables used to assign clarification forms to annotators.

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

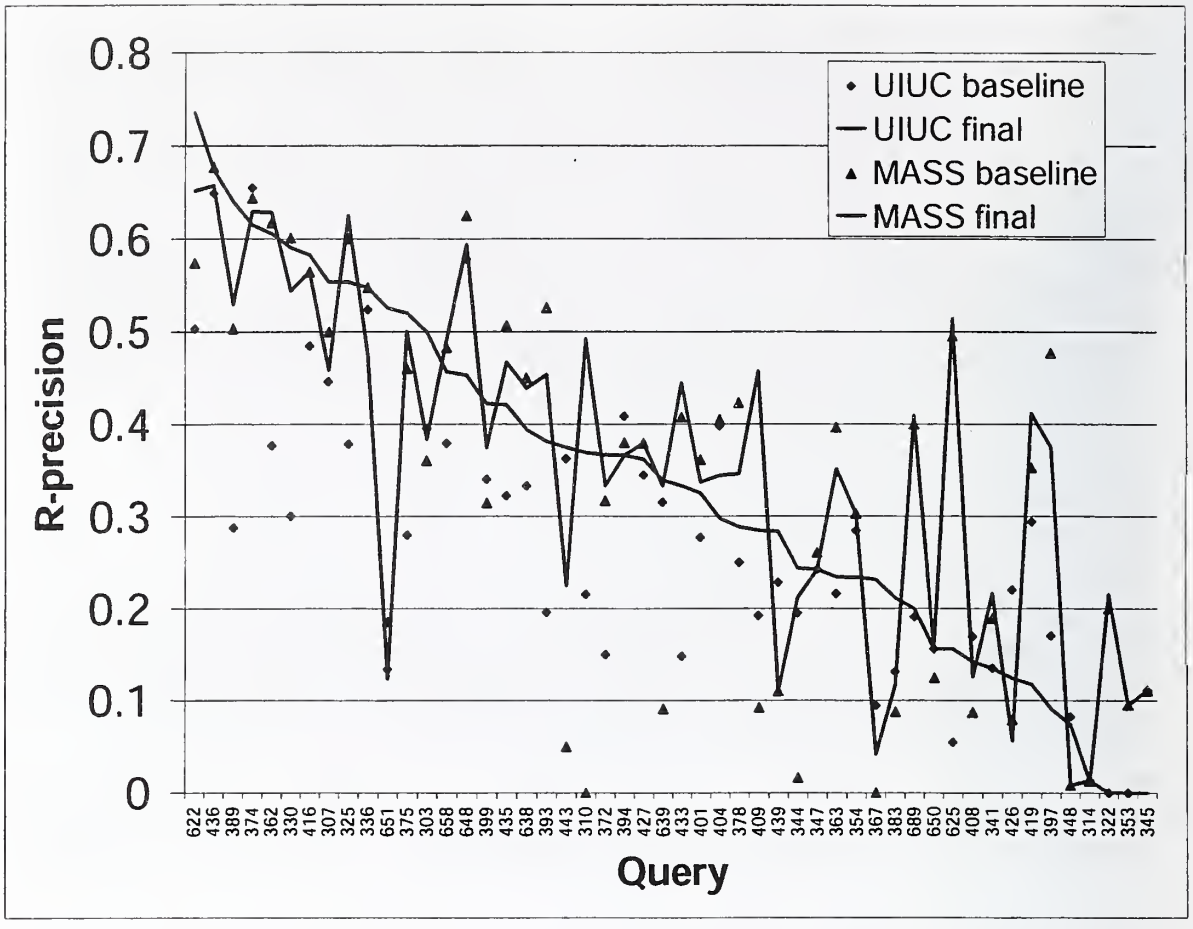

Figure 7: A query-by-query comparison of baseline and final R-precision values for the MASStrmS and UIUChCFB3 runs in Figure 6. Each query is a point on the x-axis; the queries are ordered by the final R-precision score of the UIUChCFB3 run.

References

[Allan, 2004] Allan, J. (2004). HARD track overview in TREC 2003: High accuracy retrieval from documents. In *Proceedings of TREC 2003*, pages 24–37. NIST special publication 500-255. Also on-line at http://trec.nist.gov.

[Allan, 2005] Allan, J. (2005). HARD track overview in TREC 2004: High accuracy retrieval from documents. In *Proceedings of TREC 2004*, pages 25–35. NIST special publication 500-261. Also on-line at http://trec.nist.gov.

[Baillie et al., 2006] Baillie, M., Elsweiler, D., Nicol, E., Ruthven, I., Sweeney, S., Yakici, M., Crestani, F., and Landoni, M. (2006). University of Strathclyde at TREC HARD. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Belkin et al., 2006] Belkin, N., Cole, M., Gwizdka, J., Li, Y.-L., Liu, J.-J., Muresan, G., Roussinov, D., Smith, C., Taylor, A., and Yuan, X.-J. (2006). Rutgers information interaction lab at TREC 2005: Trying HARD. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Diaz and Allan, 2006] Diaz, F. and Allan, J. (2006). When less is more: Relevance feedback falls short and term expansion succeeds at HARD 2005. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[He and Ahn, 2006] He, D. and Ahn, J. (2006). Pitt at TREC 2005: HARD and Enterprise. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Jin et al., 2006] Jin, X., French, J. C., and Michel, J. (2006). SAIC & University of Virginia at TREC 2005: HARD track. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Kelly and Fu, 2006] Kelly, D. and Fu, X. (2006). University of North Carolina's HARD track experiment at TREC 2005. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Kudo et al., 2006] Kudo, K., Imai, K., Hashimoto, M., and Takagi, T. (2006). Meiji University HARD and Robust track experiments. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Lin et al., 2006] Lin, J., Abels, E., Demner-Fushman, D., Oard, D. W., Wu, P., and Wu, Y. (2006). A menagerie of tracks at maryland: HARD, enterprise, QA, and genomics, oh my! In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Lv and Zhao, 2006] Lv, B. and Zhao, J. (2006). NLPR at TREC 2005: HARD experiments. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Rode et al., 2006] Rode, H., Ramírez, G., Westerveld, T., Hiemstra, D., and de Vries, A. P. (2006). The Lowlands' TREC experiments 2005. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Tan et al., 2006] Tan, B., Velivelli, A., Fang, H., and Zhai, C. (2006). Interactive construction of query language models – UIUC TREC 2005 HARD track experiments. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Vechtomova et al., 2006] Vechtomova, O., Kolla, M., and Karamuftuoglu, M. (2006). Experiments for HARD and Enterprise tracks. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Voorhees, 2006] Voorhees, E. M. (2006). Overview of the TREC 2005 robust retrieval track. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Wen et al., 2006] Wen, M., Huang, X., An, A., and Huang, Y. (2006). York University at TREC 2005: HARD track. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Yang et al., 2006] Yang, K., Yu, N., George, N., Loehrlen, A., McCaulay, D., Zhang, H., Akram, S., Mei, J., and Record, I. (2006). WIDIT in TREC 2005 HARD, Robust, and SPAM tracks. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Zhang et al., 2006] Zhang, J., Sun, L., Lv, Y., and Zhang, W. (2006). Relevance feedback by exploring the different feedback sources and collection structure. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

Overview of the TREC 2005 Question Answering Track

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

Abstract

The TREC 2005 Question Answering (QA) track contained three tasks: the main question answering task, the document ranking task, and the relationship task. In the main task, question series were used to define a set of targets. Each series was about a single target and contained factoid and list questions. The final question in the series was an "Other" question that asked for additional information about the target that was not covered by previous questions in the series. The main task was the same as the single TREC 2004 QA task, except that targets could also be events; the addition of events and dependencies between questions in a series made the task more difficult and resulted in lower evaluation scores than in 2004. The document ranking task was to return a ranked list of documents for each question from a subset of the questions in the main task, where the documents were thought to contain an answer to the question. In the relationship task, systems were given TREC-like topic statements that ended with a question asking for evidence for a particular relationship.

The goal of the TREC question answering (QA) track is to foster research on systems that return answers themselves, rather than documents containing answers, in response to a question. The track started in TREC-8 (1999), with the first several editions of the track focused on *factoid* questions. A factoid question is a fact-based, short answer question such as *How many calories are there in a Big Mac?*. The task in the TREC 2003 QA track contained list and definition questions in addition to factoid questions [1]. A list question asks for different instances of a particular kind of information to be returned, such as *List the names of chewing gums*. Answering such questions requires a system to assemble an answer from information located in multiple documents. A definition question asks for interesting information about a particular person or thing such as *Who is Vlad the Impaler?* or *What is a golden parachute?*. Definition questions also require systems to locate information in multiple documents, but in this case the information of interest is much less crisply delineated.

In TREC 2004 [2], 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.

The TREC 2005 QA track contained three tasks: the main question answering task, the document ranking task, and the relationship task. The document collection from which answers were to be drawn was the AQUAINT Corpus of English News Text (LDC catalog number LDC2002T31). The main task was the same as the TREC 2004 task, with one significant change: in addition to persons, organizations, and things, the target could also be an event. Events were added in response to suggestions that the question series include answers that could not be readily found by simply looking up the target in Wikipedia or other pre-compiled Web resources. The runs were evaluated using the same methodology as in TREC 2004, except that the primary measure was the per-series score instead of the combined component score.

The document ranking task was added to build infrastructure that would allow a closer examination of the role document retrieval techniques play in supporting QA technology. The task was to submit, for a subset of 50 of the questions in the main task, a ranked list of up to 1000 documents for each question. The purpose of the lists was to create document pools both to get a better understanding of the number of instances of correct answers in the collection and to support research on whether some document retrieval techniques are better than others in support of QA. NIST

pooled the document lists for each question, and assessors judged each document in the pool as relevant ("contains an answer") or not relevant ("does not contain an answer"). Document lists were then evaluated using trec_eval measures.

Finally, the relationship task was added. The task was the same as was performed in the AQUAINT 2004 relationship pilot. Systems were given TREC-like topic statements that ended with a question asking for evidence for a particular relationship. The initial part of the topic statement set the context for the question. The question was either a yes/no question, which was understood to be a request for evidence supporting the answer, or an explicit request for the evidence itself. The system response was a set of information nuggets that were evaluated using the same scheme as definition and Other questions.

The remainder of this paper describes each of the three tasks in the TREC 2005 QA track in more detail. Section 1 describes the question series that formed the basis of the main and document ranking tasks; section 2 describes the evaluation method and resulting scores for the runs for the main task, while section 3 describes the evaluation and results of the document ranking task. The questions and results for the relationship task are described in section 4. Section 5 summarizes the technical approaches used by the systems to answer the questions, and the final section looks at the future of the track.

1 Question Series

The main task for the TREC 2005 QA track required providing answers for each question in a set of question series. A question series consists of several factoid questions, one to two list questions, and exactly one Other question. Associated with each series is a definition target. The series that a question belongs to, the order of the question in the series, and the type of each question (factoid, list, or Other) are all explicitly encoded in the XML format used to describe the test set. Example series (minus the XML tags) are shown in figure 1.

95	return of Hong Kong to Chinese sovereignty		
	95.1	FACTOID	What is Hong Kong's population?
	95.2	FACTOID	When was Hong Kong returned to Chinese sovereignty?
	95.3	FACTOID	Who was the Chinese President at the time of the return?
	95.4	FACTOID	Who was the British Foreign Secretary at the time?
	95.5	LIST	What other countries formally congratulated China on the return?
	95.6	OTHER	
111	AMWAY		
	111.1	FACTOID	When was AMWAY founded?
	111.2	FACTOID	Where is it headquartered?
	111.3	FACTOID	Who is the president of the company?
	111.4	LIST	Name the officials of the company.
	111.5	FACTOID	What is the name "AMWAY" short for?
	111.6	OTHER	
136	Shiite		
	136.1	FACTOID	Who was the first Imam of the Shiite sect of Islam?
	136.2	FACTOID	Where is his tomb?
	136.3	FACTOID	What was this person's relationship to the Prophet Mohammad?
	136.4	FACTOID	Who was the third Imam of Shiite Muslims?
	136.5	FACTOID	When did he die?
	136.6	FACTOID	What portion of Muslims are Shiite?
	136.7	LIST	What Shiite leaders were killed in Pakistan?
	136.8	OTHER	

Figure 1: Sample question series from the test set. Series 95 has an EVENT as a target, series 111 has an ORGANIZATION as a target, and series 136 has a THING as a target.

The scenario for the main task was that an adult, native speaker of English was looking for more information about

a target that interested him. The target could be a person, organization, thing, or event. The user was assumed to be an "average" reader of U.S. newspapers. NIST assessors acted as surrogate users and developed the question and judged the system responses.

In TREC 2004, the question series had been written primarily *before* the assessors had searched the AQUAINT document collection; consequently, many of the question series had been unusable because the document collection did not have sufficient information to answer the questions. Therefore, the questions for TREC 2005 were developed by the assessors *after* searching the AQUAINT document collection to make sure that there was sufficient information about the target. The assessors created factoid and list questions whose answers could be found in the document collection; they tried to phrase the questions as something they would have asked if they hadn't seen the documents already. The assessors also recorded other interesting information that was not an answer to a factoid or list question (because the information was not a factoid, or the question would be too obviously a back-formulation of the answer), which could be used to answer the final "Other" question in the series.

Context processing is an important element for question answering systems to possess, so a question in the series could refer to the target or a previous answer using a pronoun, definite noun phrase or other referring expression, as shown in figure 1. Each series is an abstraction of an information dialogue in which the user is trying to define the target, but it is only a limited abstraction. Unlike in a real dialogue, questions could not mention (by name) an answer to a previous question in the series. Because each usable series was *required* to contain a list question whose answers were named entities, assessors sometimes asked list questions that they were not actually interested in. This means that the series may not necessarily be true samples of the assessor's interests in the target.

The final test set contained 75 series; the targets of these series are given in table 1. Of the 75 targets, 19 are PERSONs, 19 are ORGANIZATIONs, 19 are THINGs, and 18 are EVENTs. The series contained a total of 362 factoid questions, 93 list questions, and 75 (one per target) Other questions. Each series contained 6-8 questions (counting the Other question), with most series containing 7 questions.

Participants were required to submit retrieval results within one week of receiving the test set. All processing of the questions was required to be strictly automatic. Systems were required to process series independently from one another, and required to process an individual series in question order. That is, systems were allowed to use questions and answers from earlier questions in a series to answer later questions in that same series, but could not "look ahead" and use later questions to help answer earlier questions. As a convenience for the track, NIST made available document rankings of the top 1000 documents per target as produced using the PRISE document retrieval system and the target as the query. Seventy-one runs from 30 participants were submitted to the main task.

2 Main Task Evaluation

The evaluation of a single run comprises the component evaluations for each of the question types, and a final average per-series score. Each of the three question types has its own response format and evaluation method. The individual component evaluations for 2005 were identical to those used in the TREC 2004 QA track. Next, a per-series score was computed for a run using a weighted 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 for a factoid question was either exactly one [*doc-id*, *answer-string*] pair or the literal string 'NIL'. Since there was no guarantee that a factoid question had an answer in the document collection, NIL was returned by the system when it believed there was no answer. Otherwise, *answer-string* was a string containing precisely an answer to the question, and *doc-id* was the id of a document in the collection that supported *answer-string* as an answer.

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

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

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

Table 1: Targets of the 75 question series.

66	Russian submarine Kursk sinks	104	1999 North American International Auto Show
67	Miss Universe 2000 crowned	105	1980 Mount St. Helens eruption
68	Port Arthur Massacre	106	1998 Baseball World Series
69	France wins World Cup in soccer	107	Chunnel
70	Plane clips cable wires in Italian resort	108	Sony Pictures Entertainment (SPE)
71	F16	109	Telefonica of Spain
72	Bollywood	110	Lions Club International
73	Viagra	111	AMWAY
74	DePauw University	112	McDonald's Corporation
75	Merck and Co.	113	Paul Newman
76	Bing Crosby	114	Jesse Ventura
77	George Foreman	115	Longwood Gardens
78	Akira Kurosawa	116	Camp David
79	Kip Kinkel school shooting	117	kudzu
80	Crash of EgyptAir Flight 990	118	U.S. Medal of Honor
81	Preakness 1998	119	Harley-Davidson
82	Howdy Doody Show	120	Rose Crumb
83	Louvre Museum	121	Rachel Carson
84	meteorites	122	Paul Revere
85	Norwegian Cruise Lines (NCL)	123	Vicente Fox
86	Sani Abacha	124	Rocky Marciano
87	Enrico Fermi	125	Enrico Caruso
88	United Parcel Service (UPS)	126	Pope Pius XII
89	Little League Baseball	127	U.S. Naval Academy
90	Virginia wine	128	OPEC
91	Cliffs Notes	129	NATO
92	Arnold Palmer	130	tsunami
93	first 2000 Bush-Gore presidential debate	131	Hindenburg disaster
94	1998 indictment and trial of Susan McDougal	132	Kim Jong Il
95	return of Hong Kong to Chinese sovereignty	133	Hurricane Mitch
96	1998 Nagano Olympic Games	134	genome
97	Counting Crows	135	Food-for-Oil Agreement
98	American Legion	136	Shiite
99	Woody Guthrie	137	Kinmen Island
100	Sammy Sosa	138	International Bureau of Universal Postal Union (UPU)
101	Michael Weiss	139	Organization of Islamic Conference (OIC)
102	Boston Big Dig	140	PBGC
103	Super Bowl XXXIV		

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

Run Tag	Submitter	Accuracy	NIL Prec	NIL Recall
lcc05	Language Computer Corp.	0.713	0.643	0.529
NUSCHUA1	National Univ. of Singapore	0.666	0.148	0.529
IBM05L3P	IBM T.J. Watson Research	0.326	0.200	0.118
ILQUA2	Univ. of Albany	0.309	0.075	0.235
Insun05QA1	Harbin Inst. of Technology	0.293	0.057	0.176
csail2	MIT	0.273	0.098	0.294
FDUQA14B	Fudan University	0.260	0.082	0.412
QACTIS05v2	National Security Agency (NSA)	0.257	0.045	0.176
mk2005qar2	Saarland University	0.235	0.071	0.353
Edin2005b	Univ. of Edinburgh	0.215	0.068	0.176

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

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

To be responsive, an answer string was required to contain appropriate units and to refer to the correct "famous" entity (e.g., the Taj Mahal casino is not responsive when the question asks about "the Taj Mahal"). 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 are correct only if there is no known answer to the question in the collection and are incorrect otherwise. NIL is correct for 17 of the 362 factoid questions in the test set. (Eighteen questions had no correct response returned by the systems, but did have a correct answer found by the assessors.)

The main evaluation score for the factoid component is *accuracy*, the fraction of questions judged correct. Also reported are the recall and precision of recognizing when no answer exists in the document collection. NIL precision is the ratio of the number of times NIL was returned and correct to the number of times it was returned, whereas NIL recall is the ratio of the number of times NIL was returned and correct to the number of times it was correct (17). If NIL was never returned, NIL precision is undefined and NIL recall is 0.0. Table 2 gives evaluation results for the factoid component. The table shows the most accurate run for the factoid component for each of the top 10 groups. The table gives the accuracy score over the entire set of factoid questions as well as NIL precision and recall scores.

2.2 List questions

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

The final set of correct answers for a list question was compiled from the union of the correct responses across all runs plus the instances the assessor found during question development. For the 93 list questions used in the evaluation, the average number of answers per question is 12.5, with 2 as the smallest number of answers, and 70 as the maximum number of answers. A system's response to a list question was scored using instance precision (IP) and instance recall (IR) based on the list of known instances. Let S be the the number of known instances, D be the number of correct, distinct responses returned by the system, and N be the total number of responses returned by the

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

Run Tag	Submitter	F
lcc05	Language Computer Corp.	0.468
NUSCHUA3	National Univ. of Singapore	0.331
IBM05C3PD	IBM T.J. Watson Research	0.131
ILQUA1	Univ. of Albany	0.120
csail1	MIT	0.110
QACTIS05v1	National Security Agency (NSA)	0.105
Insun05QA1	Harbin Inst. of Technology	0.085
Edin2005a	Univ. of Edinburgh	0.081
MITRE2005B	Mitre Corp.	0.080
shef05lmg	Univ. of Sheffield	0.076

system. Then $IP = D/N$ and $IR = D/S$. Precision and recall were then combined using the F measure with equal weight given to recall and precision:

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

The score for the list component of a run was the average F score over the 93 questions. Table 3 gives the average F scores for 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 same methodology as the TREC 2003 definition questions. A system's response for an Other question was an unordered set of [*doc-id*, *answer-string*] pairs as in the list component. Each string was presumed to be a facet in the definition of the series' target that had not yet been covered by earlier questions in the series. The requirement to not repeat information already covered by earlier questions in the series made answering Other questions somewhat more difficult than answering TREC 2003 definition questions.

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

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

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

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

Run Tag	Submitter	F($\beta = 3$)
QACTIS05v3	National Security Agency (NSA)	0.248
FDUQA14B	Fudan University	0.232
lcc05	Language Computer Corp.	0.228
MITRE2005B	Mitre Corp.	0.217
NUSCHUA3	National Univ. of Singapore	0.211
ILQUA2	Univ. of Albany	0.207
IBM05C3PD	IBM T.J. Watson Research	0.206
uams05be3	Univ. of Amsterdam	0.201
SUNYSB05qa2	SUNY Stony Brook	0.196
UNTQA0501	Univ. of North Texas	0.191

computed as the F measure with nugget recall three times as important as nugget precision:

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

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

As a separate experiment, the University of Maryland created a manual "run" for the Other questions, in which a human wrote down what he thought were good nuggets for each of the questions. This manual run was included in the judging of the submitted automatic runs, and received an average F($\beta = 3$) score of 0.299. The low score may indicate the level of variation between humans regarding what information is considered interesting (vital or okay) for a target. However, this score should not be taken as an upper bound on system performance, since the manual run sometimes included information from previous questions in the series (which were explicitly excluded from the desired Other information). The run also had shorter answer strings than the best system responses; this resulted in high average precision (0.482) at the cost of lower recall (0.296), while the scoring method gave greater importance to recall than precision.

2.4 Per-series Combined Weighted Scores

The three component scores measure systems' 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, evaluating over the individual series 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 scores as the final score for the run. The weighted average of the three component scores for a series for a QA run is computed as:

$$\text{WeightedScore} = .5 \times \text{Factoid} + .25 \times \text{List} + .25 \times \text{Other}.$$

To compute the weighted score for an individual series, only the scores for questions belonging to the series were part of the computation. Since each of the component scores ranges between 0 and 1, the weighted score is also in that range. The average per-series weighted score is called the per-series score and gives equal weight to each series. Table 5 shows the per-series score for the best run for each of the top 10 groups.

Each individual series has only a few questions, so the combined weighted score for an individual series will be much less stable than the global score. But the average of 75 per-series scores should be at least as stable as the overall combined weighted average and has some additional advantages. The per-series score is computed at a small enough

Table 5: Per-series scores for QA task runs. Scores are given for the best run from the top 10 groups.

Run Tag	Submitter	Per-series Score
lcc05	Language Computer Corp.	0.534
NUSCHUA3	National Univ. of Singapore	0.464
IBM05C3PD	IBM T.J. Watson Research	0.246
ILQUA2	Univ. of Albany	0.241
QACTIS05v3	National Security Agency (NSA)	0.222
FDUQA14B	Fudan University	0.205
csail2	MIT	0.201
Insun05QA1	Harbin Inst. of Technology	0.187
shef05lmg	Univ. of Sheffield	0.165
mk2005qar2	Saarland University	0.158

granularity to be meaningful at the task-level (i.e., each series representing a single user interaction), and at a large enough granularity for individual scores to be meaningful. As pointed out in [2], many individual questions have zero for a median score over all runs, but only a few series have a zero median per-series score.

We fit a two-way analysis of variance model with the target type and the best run from each group as factors, and the per-series combined score as the dependent variable. Both main effects are significant at a p value essentially equal to 0, which indicates that there are significant differences between runs as well as between target types. 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 6 shows the results of the multiple comparison; runs sharing a common letter are not significantly different.

A similar analysis showed that PERSON and ORGANIZATION type targets having significantly higher per-series scores than EVENT and THING targets. System effectiveness may be higher for persons and organizations because the types of information desired for a person or organization may be more standard than for an event or thing. While it may be possible to come up with templates for events, identifying references to a particular event in a document collection is difficult because events are usually unnamed and the exent of the event is not always well-defined.

3 Document Ranking Task

The goal of the document ranking task was to create pools of documents containing answers to questions in the main series. These pools would provide an estimate of the number of instances of correct answers in the collections for people wanting to use the 2005 evaluated data for post-conference experiments. The task would also support research on whether some document retrieval techniques are better than others in support of QA, since groups were allowed to mix and match different techniques for retrieval and QA.

All TREC 2005 submissions to the main task were required to include a ranked list of documents for each question in the document ranking task; the list represented the set of documents used by the system to create its answer, where the order of the documents in the list was the order in which the system considered the document. There were 77 submissions to the document ranking task. Groups whose primary emphasis was document retrieval rather than QA, were allowed to participate in the document ranking task without submitting actual answers for the main task; three groups participated in the document ranking task without participating in the main task.

The test set for the document ranking task was a list of question numbers for 50 of the questions from the main task. The set of 50 questions comprised all the factoid and list questions from two series and additional factoid questions from other series. Half of these questions contained pronouns or other anaphors that referred to the target or answer to a previous question. For each question, systems returned a ranked list of up to 1000 documents that were thought to contain an answer for the question.

RunID	PMM													
lcc05	0.5343	A												
NUSCHUA3	0.4641		B											
IBM05C3PD	0.2457			C										
ILQUA2	0.2412			C										
QACTIS05v3	0.2219			C	D									
FDUQA14B	0.2050			C	D									
csail2	0.2004			C	D	E								
Insun05QA1	0.1868			C	D	E	F							
shef05lmg	0.1644				D	E	F	G						
mk2005qar2	0.1578				D	E	F	G						
asked05c	0.1568				D	E	F	G						
Edin2005c	0.1552				D	E	F	G	H					
clr05	0.1357					E	F	G	H	I				
UNTQA0503	0.1337					E	F	G	H	I	J			
ASUQA02	0.1332					E	F	G	H	I	J			
MITRE2005B	0.1328					E	F	G	H	I	J			
uams05be3	0.1268						F	G	H	I	J			
talpupc05b	0.1253						F	G	H	I	J	K		
SUNYSB05qa3	0.1232						F	G	H	I	J	K		
DLT05QA01	0.1183						F	G	H	I	J	K	L	
CMUJAV2005	0.1060							G	H	I	J	K	L	
Dal05s	0.0872								H	I	J	K	L	M
lexicloneB	0.0841									I	J	K	L	M
TWQA0502	0.0748									I	J	K	L	M
Mon05BIMP2	0.0699									I	J	K	L	M
thuir051	0.0654										J	K	L	M
dggQA05X	0.0568											K	L	M
MSRCOMB05	0.0542												L	M
UIowaQA0503	0.0271													M
afrun1	0.0152													

Table 6: Multiple comparison of best run from each group, based on ANOVA of per-series score. PMM is the population marginal mean of the per-series score for the run.

Table 7: R-Precision and MAP scores for the document-ranking task runs. Scores are given for the best run from the top 13 groups.

Run Tag	Submitter	R-Prec	MAP
NUSCHUA1	National Univ. of Singapore	0.4570	0.4698
* humQ05xle	Hummingbird	0.4127	0.4468
IBM05C3PD	IBM T.J. Watson Research	0.3978	0.4038
QACTIS05v1	National Security Agency (NSA)	0.3414	0.3498
* apl05aug	Johns Hopkins Univ. Applied Physics Lab	0.3201	0.3417
ASUQA01	Arizona State Univ.	0.2958	0.3321
UNTQA0501	Univ. of North Texas	0.3205	0.3285
* sab05qa1b	Sabir Research	0.3366	0.3197
lcc05	Language Computer Corp.	0.2921	0.3045
afrun1	Macquarie Univ.	0.3038	0.2852
TWQA0501	Peking Univ.	0.2732	0.2832
csail2	MIT	0.2699	0.2808
ILQUA1	Univ. of Albany	0.2445	0.2596

3.1 Evaluation

For each of the 50 questions, the documents in the top 75 ranks for up to two runs per group were pooled and then judged by the human assessor. A document was considered relevant if the document contained a correct, supported answer and not relevant otherwise. Each pool had an average of about 717 documents; the smallest pool had 295 documents, and the largest pool had 1219 documents. The number of relevant documents (containing an answer) in each pool ranged from 1 to 285, with a mean of 31.5 documents and a median of 7 documents. As expected, the number of different documents containing an answer for each question, as judged in the document ranking task, was far higher than the number of different documents containing the right answer as judged in the strict question answering task. Researchers doing post-evaluation analysis should therefore not assume that the set of documents having correct answers in the main series task is complete.

The submitted runs were scored using trec_eval, treating the contains-answer documents as the relevant documents. Unlike other QA evaluations, trec_eval rewards recall, so retrieving more documents with the same answer yields a higher score than retrieving a single document with that answer. Even though a factoid question requires only a single document containing an answer, a recall-based metric for document retrieval may still correlate with performance on the exact factoid QA task because some systems make use of the frequency of candidate answers in determining which candidate to select as the final answer.

Table 7 shows the R-Precision and mean average precision (MAP) scores for the best run for each of the top 13 groups. The runs for the three groups that participated in the document ranking task without participating in the main task are marked with a $*$. R-precision is the precision after retrieving the first R documents, where R is the number of relevant documents in the pool. We found a weak correlation between factoid accuracy and R-precision (Pearson's $\rho = 0.53$, with a 95% confidence interval of [0.38,1.0]).

4 Relationship Task

AQUAINT analysts defined a "relationship" as the ability of one entity to influence another, including both the means to influence and the motivation for doing so. Eight spheres of influence have been noted including financial, movement of goods, family ties, communication pathways, organizational ties, co-location, common interests, and temporal. Recognition of when support for a suspected tie is lacking and determining whether the lack is because the tie doesn't exist or is being hidden/missed is a major concern. The analyst needs sufficient information to establish confidence in any support given. The particular relationships of interest depend on the context.

In the relationship task, 4 military analysts created 25 TREC-like topic statements that set a context. Each topic

Figure 2: Sample relationship topic and nuggets of evidence.

The analyst is concerned with arms trafficking to Colombian insurgents. Specifically, the analyst would like to know of the different routes used for arms entering Colombia and the entities involved.	
Vital?	Nugget
vital	Weapons are flown from Jordan to Peru and air dropped over southern Columbia
okay	Jordan denied that it was involved in smuggling arms to Columbian guerrillas
vital	Jordan contends that a Peruvian general purchased the rifles and arranged to have them shipped to Columbia via the Amazon River.
okay	Peru claims there is no such general
vital	FARC receives arms shipments from various points including Ecuador and the Pacific and Atlantic coasts.
okay	Entry of arms to Columbia comes from different borders, not only Peru

Table 8: Average F($\beta = 3$) scores for the relationship task for each run. Manual runs are marked with a $*$.

Run Tag	Submitter	F($\beta = 3$)
* clr05r1	CL Research	0.276
csail2005a	MIT	0.228
* clr05r2	CL Research	0.216
* lcc05rel1	Language Computer Corp.	0.190
* lcc05rel2	Language Computer Corp.	0.171
uams05s	Univ. of Amsterdam	0.120
uams05l	Univ. of Amsterdam	0.119
* CMUJAVSEMMAN	Carnegie Mellon Univ.	0.096
* UIowa05QAR01	Univ. of Iowa	0.086
CMUJAVSEM	Carnegie Mellon Univ.	0.061

was specific about the type of relationship being sought. The topic ended with a question that was either a yes/no question, which was to be understood as a request for evidence supporting the answer, or a request for the evidence itself. The system response was a set of information nuggets that provided evidence for the answer, in the same format as the Other questions in the main task. Manual processing was allowed.

4.1 Evaluation

The relationship topics were evaluated using the same methodology as the Other questions in the main task. A system's response for a relationship topic was an unordered set of [*doc-id*, *answer-string*] pairs. Each string was presumed to contain evidence for the answer to the question(s) in the topic. The system responses were judged by 5 assessors who were not the same as those who created the topics. An example topic and associated nuggets of evidence are given in Figure 2.

Each nugget created by the assessor was a piece of evidence for the answer, with nuggets marked as either vital or non-vital. Precision, recall, and F measure were calculated for each relationship topic as for the Other questions, and the final score for the relationship task was the average F($\beta = 3$) score over 25 topics. Table 8 gives the average F($\beta = 3$) score for each of the 10 runs submitted for the relationship task. Runs that included manual processing are marked with a $*$.

5 System Approaches

The overall approach taken for answering factoid questions has remained unchanged for the past several years. Systems generally determine the expected answer type of the question, retrieve documents or passages likely to contain answers to the question using important question words and related terms as the query, and then perform a match between the question words and retrieved passages to generate a set of candidate answers. The candidate answers are then ranked to find the most likely answer.

For the document/passage retrieval phase, most systems simply appended the target to the query. This was an effective strategy since in all cases the target was the correct domain for the question, and most of the retrieval methods used treat the query as a simple set of keywords. More and more systems are exploiting the size and redundancy on the Web to help find the answer. Some search the Web to find the answer, and then project the answer back to the AQUAINT corpus to find a supporting document. Others find candidate answers in the AQUAINT corpus and then use the Web to rerank the answers.

Most groups use their factoid-answering system for list questions, returning the top-ranked n candidate answer strings as the final answer list. The number of answer strings returned was a fixed number or was based on some threshold score for the string. Some groups went further and used their initial list items as seeds to find additional items. Systems generally used the same techniques as were used for TREC 2003's definition questions to answer the Other and relationship questions. Most systems first retrieve passages about the target using a recall-oriented retrieval search. Subsequent processing reduces the amount of material returned. Systems also looked to eliminate redundant information, using either word overlap measures or document summarization techniques. The output from the redundancy-reducing step was then returned as the answer for the question.

6 Future of the QA Track

Even though the main task in the TREC 2005 QA task was supposed to be essentially the same as the 2004 task, system performance was noticably lower in 2005 than in 2004. The 2005 task was more difficult because of the introduction of EVENT type targets and the increased dependencies between questions in a series; questions contained a greater number of anaphoric references, many of which referred to answers to previous questions in the series.

The introduction of event targets had additional ramifications for NIST assessors judging the system responses; it became clear that the assessors would not (and should not) ignore the time frame implied by the series when judging the correctness of answers. Before 2005, assessors assumed that the document returned with an answer would be used to set the time frame for the question, because questions were primarily phrased in the present tense without specifying an explicit time frame. Under those guidelines, *Who is the President of the United States?* would be answered correctly by "Ronald Reagan" if the document was from 1987, even if more recent documents supported "George Bush" or "Bill Clinton" as the answer. However, event type targets and temporally-constrained questions require that questions be interpreted in the temporal context that is explicit in the question or implicit in the series.

The main task for the TREC 2006 QA track will be the same as the main task in 2005, except that the implicit time frame for questions phrased in the present tense will be the date of the last document in the document collection, rather than the document returned with the answer. Thus, systems will be required to give the most up-to-date answer supported by the document collection. This brings the TREC QA task closer in line with question-answering in the real world, where users would want the best answer to their question in the document set (rather than just any answer found in any document). The evaluation of the question series in 2006 will also weight each of the 3 question types equally. The document ranking task will not be repeated in 2006, since little was learned from it. However, the relationship task will be repeated and modified to allow clarification forms like the ones used in the 2005 HARD task.

References

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

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

Overview of the TREC 2005 Robust Retrieval Track

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

Abstract

The robust retrieval track explores methods for improving the consistency of retrieval technology by focusing on poorly performing topics. The retrieval task in the track is a traditional ad hoc retrieval task where the evaluation methodology emphasizes a system's least effective topics.

The 2005 edition of the track used 50 topics that had been demonstrated to be difficult on one document collection, and ran those topics on a different document collection. Relevance information from the first collection could be exploited in producing a query for the second collection, if desired. The main measure for evaluating system effectiveness is "gmap", a variant of the traditional MAP measure that uses a geometric mean rather than an arithmetic mean to average individual topic results. As in previous years, the most effective retrieval strategy was to expand queries using terms derived from additional corpora. The relative difficulty of topics differed across the two document sets.

Systems were also required to rank the topics by predicted difficulty. This task is motivated by the hope that systems will eventually be able to use such predictions to do topic-specific processing. This remains a challenging task. Since difficulty depends on more then the topic set alone, prediction methods that train on data from other test collections do not generalize well.

The ability to return at least passable results for any topic is an important feature of an operational retrieval system. While system effectiveness is generally reported as average effectiveness, an individual user does not see the average performance of the system, but only the effectiveness of the system on his or her request. The previous two editions of the robust track have demonstrated that average effectiveness masks individual topic effectiveness, and that optimizing standard average effectiveness measures usually harms the already ineffective topics.

This year's track used 50 topics that had been demonstrated to be difficult for the TREC Disks 4&5 document set (CD45) and ran those topics against the AQUAINT document set. Relevance information from the CD45 collection could be exploited in producing a query for the AQUAINT collection, if desired.

A focus of the robust track since its inception has been developing the evaluation methodology for measuring how well systems avoid abysmal results for individual topics. Two measures introduced in the initial track were subsequently shown to be relatively unstable even for as many as 100 topics in the test set [3]. Those measures have been dropped from this year's results and have been replaced by the geometric MAP, or "gmap", measure. Gmap is computed as a geometric mean of the average precision scores of the test set of topics, as opposed to the arithmetic mean used to compute the standard MAP measure. Experiments using the TREC 2004 robust track results suggest that the measure gives appropriate emphasis to poorly performing topics while being stable with as few as 50 topics.

In addition to producing a ranked list of documents for each topic, systems were also required to rank the topics by predicted difficulty. The motivation for this task is the hope that systems will eventually be able to use such predictions to do topic-specific processing.

This paper presents an overview of the results of the track. The first section describes the data used in the track, and the following section gives the systems' retrieval results. Section 3 examines the differences in the test collections built with the different document sets. Despite the diversity of runs that contributed to the pools for the AQUAINT collection, analysis of the resulting relevance judgments suggests the pool depth was insufficient with respect to the document set size. Section 4 then examines the difficulty prediction task. The final section summarizes the results of the three-year run of the track: this is the concluding year of a separate robust track, though the gmap measure with its emphasis on poorly performing topics will be incorporated into ad hoc tasks in other tracks.

1 The Robust Retrieval Task

The task within the robust retrieval track is a traditional ad hoc task. The document set used in this year's track was the AQUAINT Corpus of English News Text (LDC catalog number LDC2002T31). This collection consists of documents from three different sources: the AP newswire from 1998–2000, the New York Times newswire from 1998–2000, and the (English portion of the) Xinhua News Agency from 1996–2000. There are approximately 1,033,000 documents and 3 gigabytes of text in the collection.

The topic set consisted of 50 topics that had been used in ad hoc and robust tracks in previous years where they were run against the document set comprised of the documents on TREC disks 4&5 (minus the *Congressional Record*). These topics each had low median average precision scores in both the initial TREC in which they were used and in previous robust tracks, and were chosen for the track precisely because they are assumed to be difficult topics.

The 50 test topics were selected from a somewhat larger set based on having at least three relevant documents in the AQUAINT collection. NIST assessors were given a set of of topic statements and asked to search the AQUAINT collection looking for at least three relevant documents. Assessors were given the general guideline that they should spend no more than about 30 minutes searching for any one topic. The assessor stopped searching for relevant documents as soon as he or she found three relevant documents or when they felt they had exhausted the collection without finding three relevant documents. The topics for which fewer than three relevant documents were retrieved were discarded. The entire process stopped as soon as 50 topics with a minimum of three relevant documents were found.

The assessor who judged a topic on the AQUAINT data set was in general different from the assessor who originally judged the topic on the CD45 collection. Thus, both the document set and the assessor differed between original runs using the topics and the robust 2005 runs. Nonetheless, systems were allowed to exploit the existing judgments in creating their queries for the track if they chose to do so. (Such runs were labeled as manual or "human-assisted" runs since the previous judgments were manually created. Runs that used other types of manual processing are also labeled as human-assisted.) Using the existing judgments in this manner is equivalent to the routing task performed in early TRECs.

The TREC 2005 HARD track used the same test collections as the robust track. Pools for document judging were created from one baseline and one final run for each HARD track participant, and one run per robust track participant. Because there were limited assessing resources, relatively shallow pools were created. The top 55 documents per topic for each pool run were added to the pools, producing pools that had a mean size of 756 documents (minimum 350, maximum 1390). While these pools are shallow, the expectation was that the diversity of the runs used to make the pools would result in sufficiently comprehensive relevance judgments. This hypothesis is explored later in section 3. Documents in the pools were judged not relevant, relevant, or highly relevant, with both highly relevant and relevant judgments used as the relevant set for evaluation.

Runs were evaluated using `trec_eval`, and the standard measures are included in the evaluation report for robust runs. The primary measure for the track is the geometric MAP (gmap) score computed over the 50 test topics. Gmap was introduced in the TREC 2004 robust track [3] as a measure that emphasizes poorly performing topics while remaining stable with as few as 50 topics. Gmap takes a geometric mean of the individual topics' average precision scores, which has the effect of emphasizing scores close to 0.0 (the poor performers) while minimizing differences between larger scores. The geometric mean is equivalent to taking the log of the the individual topics' average precision scores, computing the arithmetic mean of the logs, and exponentiating back for the final gmap score. The gmap value reported for robust track runs was computed using the current version of `trec_eval` (invoked with the -a option). In this implementation, all individual topic average precision scores that are less than 0.00001 are set to 0.00001 to avoid taking logs of 0.0.

2 Retrieval Results

The robust track received a total of 74 runs from the 17 groups listed in Table 1. Participants were allowed to submit up to five runs. To have comparable runs across participating sites, if the participant submitted any automatic runs, one run was required to use just the description field of the topic statements, and one run was required to use just the title field of the topic statements. Four of the runs submitted to the track were human-assisted runs; the remaining seventy were completely automatic runs. Of the automatic runs, 24 runs were description-only runs, 34 were title-only runs,

Table 1: Groups participating in the robust track.

Arizona State University (Roussinov)	Chinese Academy of Sciences (ICT)
Ecole des Mines de Saint-Etienne	The Hong Kong Polytechnic University
Hummingbird	IBM Research, Haifa
Indiana University	IRIT/SIG
Johns Hopkins University/APL	Meiji University
Queens College, CUNY	Queensland University of Technology
RMIT University	Sabir Research, Inc.
University of Illinois at Chicago	University of Illinois at Urbana-Champaign
University of Massachusetts	

Table 2: Evaluation results for the best title-only and description-only runs for the top eight groups ordered by gmap.

Title-only Runs				Description-only Runs			
Run	gmap	MAP	P10	Run	gmap	MAP	P10
uic0501	0.233	0.310	0.592	ASUDE	0.178	0.289	0.536
indri05RdmmT	0.206	0.332	0.524	indri05RdmeD	0.161	0.282	0.498
pircRB05t2	0.196	0.280	0.542	ICT05qerfD	0.155	0.259	0.446
ICT05qerfTg	0.189	0.271	0.444	JuruDWE	0.129	0.230	0.472
UIUCrAt1	0.189	0.268	0.498	pircRB05d1	0.125	0.230	0.466
JuruTiWE	0.157	0.239	0.496	sab05rod1	0.114	0.184	0.404
humR05txle	0.150	0.242	0.490	humR05dle	0.114	0.201	0.432
wdf1t3qs0	0.149	0.235	0.456	wdf1t3qd	0.110	0.187	0.376

and 12 used various combinations of the topic statement.

Table 2 gives the evaluation scores for the best run for the top eight groups who submitted either a title-only run or a description-only run. The table gives the gmap, MAP, and average P(10) scores over the 50 topics. The run shown in the table is the run with the highest gmap; the table is sorted by this same value.

As in previous robust tracks, the best performing runs used some sort of external corpus to perform query expansion. Usually the external corpus was the web as viewed from the results of web search engines, though other large data sets such as a collection of TREC news documents (University of Massachusetts) or the .GOV collection (Chinese Academy of Sciences) were used as well. The behavior of the topics on the CD45 and AQUAINT document sets (examined in more detail below) is sufficiently different that expanding queries using a large external corpus was more effective on average than exploiting relevance information from the CD45 collection. For example, IBM Hafia found that using web expansion was more effective than no expansion, but expanding based on the CD45 relevance information was less effective than no expansion [4]. Sabir Research used the CD45 relevance information to produce "optimal" queries in its `sab05ror1` run [1]; these queries produced the best average precision scores for nine topics on the AQUAINT collection, but the average effectiveness across all topics was less than that of the best performing runs.

The top title-only runs, `uic0501` from the University of Illinois at Chicago and `indri05RdmmT` from the University of Massachusetts, illustrate the difference between the gmap and MAP measures. The `uic0501` run obtained a higher gmap score than the `indri05RdmmT` run, while the reverse is true for MAP. Figure 1 shows the per-topic average precision scores for the two runs. In the figure the topics are plotted on the x-axis and are sorted by decreasing average precision score obtained by the `indri05RdmmT` run. The horizontal line in the graph is plotted at an average precision of 0.05. The `indri05RdmmT` run has a better average precision score for more topics, but has seven topics for which the average precision score is less than 0.05. In contrast, the `uic0501` run has only two topics with an average precision score less than 0.05.

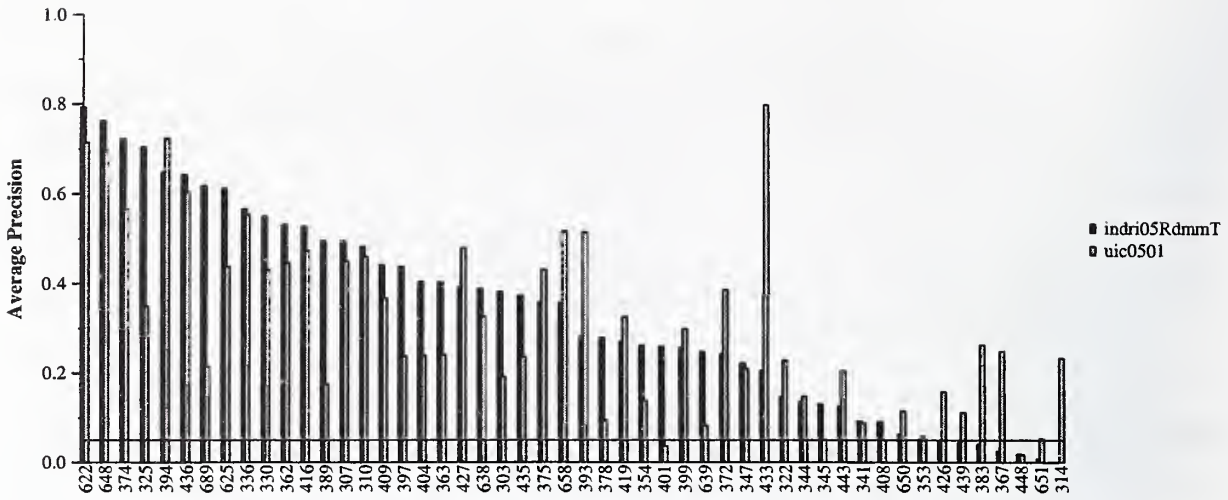

Figure 1: Per-topic average precision scores for top title-only runs. The `uic0501` run has a higher gmap score since it has fewer topics with a score less than 0.05, while the `indri05RdmmT` run has a higher average precision score for more topics and a greater MAP score.

3 The AQUAINT Test Collection

Retrieval effectiveness is in general better on the AQUAINT collection than the CD45 collection as illustrated in figure 2. The figure shows box-and-whisker plots of the average precision scores for each of the topics across the set of description-only runs submitted to TREC 2005 (top plot) and TREC 2004 (bottom plot). The line in the middle of a box indicates the median average precision score for that topic. The plots are computed over different numbers of runs (24 description-only runs in TREC 2005 vs. 30 description-only runs in TREC 2004) and in general involve different systems, but aggregate scores should be valid to compare. The majority of topics have higher medians for TREC 2005 than for TREC 2004. It is extremely unlikely that the entire set of systems that submitted description-only runs to TREC 2005 are significantly improved over TREC 2004 systems. Instead, these results remind us that topics are not inherently easy or difficult in isolation—the difficulty depends on the interaction between the information need and information source.

There are a number of differences between the ACQUAINT and CD45 test collections. The AQUAINT document set is much larger than the disks 4&5 document set: AQUAINT has more than one million documents and 3 gigabytes of text while the CD45 collection has 528,000 documents and 1904 MB of text. The AQUAINT collection contains newswire data only while the CD45 collection contains the 1994 *Federal Register* and FBIS documents. The AQUAINT collection covers a later time period. Different people assessed a given topic for the two collections. Any or all of these differences could affect retrieval effectiveness.

Earlier work in the TREC VLC track demonstrated that P(10) scores generally increase when the size of the document set increases [2]. The near doubling of the number of documents between the CD45 and AQUAINT document sets is likely a major reason for the increase in absolute scores. Aggregate statistics regarding the number of relevant documents for the two collections are not starkly different—for the AQUAINT test set there is a mean of 131.2 relevant documents per topic with a minimum number of relevant of 9 and a maximum number of relevant of 376, while the corresponding statistics for the CD45 test set are are a mean of 86.4, minimum 5, and maximum 361. But as figure 3 shows, the AQUAINT collection has many fewer topics with very small numbers of relevant documents. The figure contains a histogram of the number of relevant documents per topic for the two collections. The AQUAINT collection has only 2 topics with fewer than 20 relevant documents while the CD45 collection has 9 such topics. Good early precision scores are clearly easier to obtain when there are more relevant documents.

As figure 2 suggests, however, it is not the case that effectiveness scores simply increased by some common amount for all topics. The relative difficulty of the topics differs between the two collections. Figure 4 shows the topics sorted

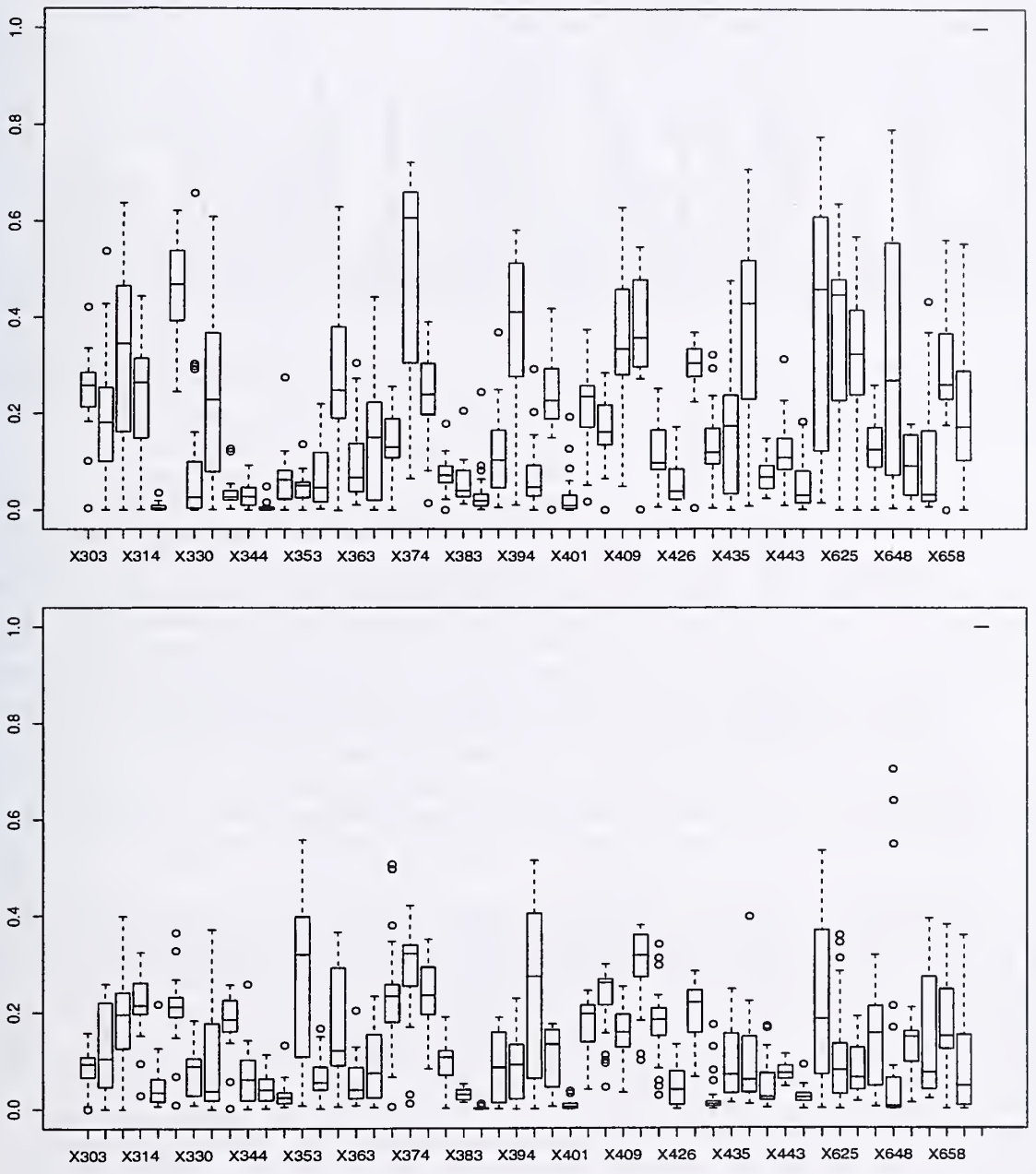

Figure 2: Box-and-whiskers plot of average precision scores for each of the 50 TREC 2005 test topics across description-only runs submitted to TREC 2005 (top) and TREC 2004 (bottom).

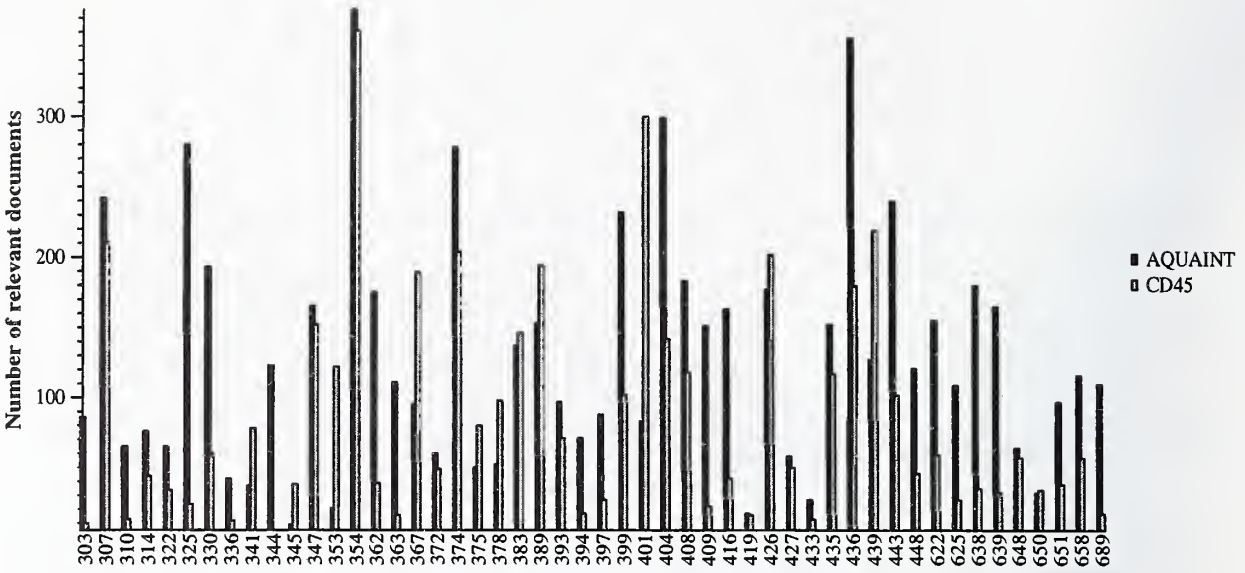

Figure 3: Number of relevant documents per topic in the TREC 2005 test set for the AQUAINT and CD45 document sets.

TREC 2005	374	325	622	625	436	394	416	310	409	638
	427	648	314	658	362	303	375	336	404	399
	435	307	689	367	408	372	639	433	443	393
	650	419	363	378	439	347	353	397	354	426
	383	651	448	344	341	330	389	401	345	322
TREC 2004	374	353	416	397	408	372	375	427	314	325
	310	404	622	341	419	639	658	409	650	399
	362	378	307	394	625	303	367	330	393	435
	651	443	638	344	354	436	689	345	336	363
	426	322	383	347	439	448	433	648	401	389
Kendall τ between rankings:			0.326							

Figure 4: Ranking of TREC 2005 test topics by decreasing median average precision score across description-only runs.

from easiest to hardest for the two collections. The difficulty of a topic is defined here as the median average precision score as computed over description-only runs submitted to either TREC 2004 and TREC 2005. The Kendall τ score between the two topic rankings is only 0.326, demonstrating that the topics have different relative difficulty on the two document sets.

The pools from which the AQUAINT test collection was created were more shallow than previous pools. Topics first used in the ad hoc tasks for TRECs 6–8 (topics 301–450) in particular had pools that were deeper and were comprised from more groups' runs than this year's pools. The expectation when the pools were formed was that the pools would be of sufficient quality because the runs contributing to the pools included both routing-type runs from the robust track and runs created after clarification form interaction from the HARD track. Unfortunately, the track results suggest that the resulting relevance judgments are dominated by a certain kind of relevant document—specifically, relevant documents that contain topic title words—and thus the AQUAINT test collection will be less reliable for future experiments where runs retrieve documents without a title-word emphasis. Note that the results of this year's HARD and robust tracks remain valid since runs from those tracks were judged.

There were two initial indications that the AQUAINT collection might be flawed. First, title-only runs are more effective than description-only runs for the AQUAINT collection, while the opposite is true for the CD45 collection. While hardly conclusive evidence of a problem, title-only queries would be expected to be better if the AQUAINT collection's shallow pools contain only easy-to-retrieve relevant documents. Second, the "optimal query" run produced by Sabir Research, a run that explicitly did not rely only on topic title words, contributed 405 unique relevant documents to the pools across the 50 topics (out of a total of $55 \times 50 = 2750$ documents contributed to the pools).

A unique relevant document is a document that was judged relevant and was contributed to the pool by exactly one group. Such documents would not have been in the pool, and therefore would be assumed irrelevant, if the one group that retrieved it had not participated in the collection building process. The difference in evaluation scores when a run is evaluated with and without the unique relevant documents from its group is used as an indication of how reusable a test collection is, since future users of the collection will not have the opportunity for their runs to be judged. The Sabir run's MAP score suffered a degradation of 23% when evaluated without its unique relevant documents, a definite warning sign.

As a result of these findings, Chris Buckley of Sabir Research and NIST examined the relevance judgments more closely. We defined a measure called titlestat as the percentage of a set of documents that a topic title word occurs in, computed as follows. For each word in the title of the current topic that is not a stop word, calculate the percentage of the set of documents, C, that contains that word, normalized by the maximum possible percentage. (The normalization is necessary because in rare cases a title word will have a collection frequency smaller than $|C|$.) Average over all title words for the topic, then average over all topics in the collection. A maximum value of 1.0 is obtained when all the documents in the set contain all topic title words; a minimum value of 0.0 means that all documents in the set contain no title words at all. Titlestat computed over the known relevant documents for the AQUAINT collection is 0.719, while the corresponding value for the CD45 collection is only 0.588. Further, the titlestat values computed over individual topics' relevant sets was greater for the AQUAINT collection than for the CD45 collection for 48 of the 50 topics.

None of the differences between the CD45 and AQUAINT document sets can plausibly explain such a change in the frequency of occurrence of topic title words in the relevant documents. If anything, title words would be expected to occur more frequently in the longer CD45 documents. Instead, the most likely explanation is that pools did not contain the documents with fewer topic title words that would have been judged relevant had they been in the pool. Topic title words are generally good descriptors of the information need stated in the topic, and retrieval systems naturally emphasize those words in their retrieval (especially when one of the mandated conditions of the track is to produce queries using only the title section!). In a collection with as many documents as the AQUAINT collection, there will be many documents containing topic title words, and these documents will fill up the pools before documents containing fewer title words will have a chance to be added.

The `sab05ror1` Sabir run further supports that contention that the majority of pool runs are dominated by documents containing topic title words while other relevant documents do exist. The titlestat computed over `sab05ror1`'s retrieved set is 0.388 while the average titlestat on the retrieved sets of the other robust track runs is 0.600. Using the unique relevant documents retrieved by the `sab05ror1` run as the set of documents the titlestat is computed over results in a value of 0.530, compared to a titlestat of 0.719 for all known relevants (including the unique relevants of the Sabir run).

Zobel demonstrated that the quality of a test collection built through pooling depends on both the diversity of the runs that contribute to the pools and the depth to which the runs are pooled [5]. In those experiments he down-sampled from existing TREC pools and saw problems only when the pools were very shallow in absolute terms. These results demonstrate how "too shallow" is relative to the document set size, a disappointing if not unexpected finding. As document collections continue to grow, traditional pooling will not be able to scale to create ever-larger reusable test collections. One of the goals of the TREC terabyte track is to examine how to build test collections for large document sets.

4 Predicting difficulty

Having a system predict whether it can effectively answer a topic is a necessary precursor to having that system modify its behavior to avoid poor performers. The difficulty prediction task was introduced into the robust track in

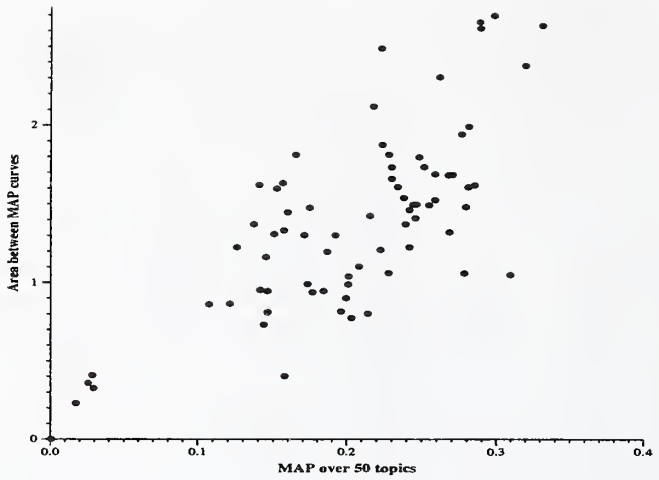

Figure 5: Scatter plot of area prediction measure vs. MAP for TREC 2005 robust track runs illustrating strong positive correlation of the scores.

TREC 2004. The task requires systems to rank the test set topics in strict order from 1 to 50 such that the topic at rank 1 is the topic the system predicted it had done best on, the topic at rank 2 is the topic the system predicted it had done next best on, etc.

Since relevance data from the CD45 collection was available for the test topics, some groups tried using that data to train difficulty predictors. These attempts were largely unsuccessful, though, since topic difficulty varied across the collections.

The difficulty-predicting task is also hampered by the lack of a suitable measure of how well a system can perform the task. Call the ranking submitted by a system its *predicted* ranking, and the topics ranked by the average precision scores obtained by the system the *actual* ranking. Clearly the quality of a system's prediction is a function of how different the predicted ranking is from the actual ranking, but this has been difficult to operationalize. The original measure used in 2004 for how the rankings differed was the Kendall τ measure between the two rankings, though it quickly became obvious that this is not a good measure for the intended goal of the predictions. The Kendall τ measure is sensitive to any change in the ranking across the entire set of topics, while the task is focused on the poor performers. A second way to measure the difference in the rankings is to look at how MAP scores change when successively greater numbers of topics are eliminated from the evaluation. In particular, compute the MAP score for a run over the best X topics where $X = 50 \ldots 25$ and the best topics are defined as the first X topics in either the predicted or actual ranking. The difference between the two curves produced using the actual ranking on the one hand and the predicted ranking on the other is the measure of how accurate the predictions are.

While the area between the two curves is a better match than Kendall τ as a quality measure of predictions for our task, it has its own faults. The biggest fault is that the area between the MAP curves is dependent on the quality of the run itself, making the area measure alone unreliable as a gauge of how good the prediction was. For example, poorly performing runs will have a small area (implying good prediction) simply because there is no room for the graphs to differ. Figure 4 shows a scatter plot of the area measure vs. the MAP score over all 50 topics for each of the runs submitted to the TREC 2005 robust track. A perfect submission would have a MAP of 1.0 and an area score of 0.0, making the lower right corner of the graph the target. Unfortunately, the strong bottom-left to top-right orientation of the plot illustrates the dependency between the two measures. Some form of normalization of the area score by the full-set MAP score may render the measure more usable.

5 Conclusion

The TREC 2005 edition of the robust retrieval track was the third, and final, running of the track in TREC. The results of the track in the various years demonstrated how optimizing average effectiveness for standard measures generally

degrades the effectiveness of poorly performing topics even further. While pseudo-relevance feedback within the target collection helps only the topics that have at least a moderate level of effectiveness to begin with, expanding queries using external corpora can be effective for poorly performing topics as well. The gmap measure introduced in the track is a stable measure that emphasizes a system's worst topics. Such an emphasis can help system builders tune their systems to avoid topics that fail completely. Gmap has been incorporated into the newest version of the trec_eval software, and will be reported for future ad hoc tasks in TREC.

References

[1] Chris Buckley. Looking at limits and tradeoffs: Sabir Research at TREC 2005. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, 2006. `http://trec.nist.gov/pubs/trec14/papers/sabir.tera.robust.qa.pdf`.

[2] David Hawking and Stephen E. Robertson. On collection size and retrieval effectiveness. *Information Retrieval*, 6(1):99–105, 2003.

[3] Ellen M. Voorhees. Overview of the TREC 2004 robust retrieval track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004*, pages 70–79, 2005.

[4] Elad Yom-Tov, David Carmel, Adam Darlow, Dan Pelleg, Shai Errera-Yaakov, and Shai Fine. Juru at TREC 2005: Query prediction in the terabyte and the robust tracks. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, 2006. `http://trec.nist.gov/pubs/trec14/papers/ibm-haifa.tera.robust.pdf`.

[5] 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.

TREC 2005 Spam Track Overview

Gordon Cormack and Thomas Lynam
University of Waterloo
Waterloo, Ontario, Canada

Abstract

TREC's *Spam Track* introduces a standard testing framework that presents a chronological sequence of email messages, one at a time, to a spam filter for classification. The filter 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. The gold standard for each message is communicated to the filter immediately following classification. Eight test corpora – email messages plus gold standard judgements – were used to evaluate 53 subject filters. Five of the corpora (the *public* corpora) were distributed to participants, who ran their filters on the corpora using a track-supplied toolkit implementing the framework. Three 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. Twelve groups participated in the track, submitting 44 filters for evaluation. The other nine subject filters were variants of popular open-source implementations adapted for use in the toolkit in consultation with their authors.

1 Introduction

The spam track's purpose is to model an email spam filter's usage as closely as possible, to measure quantities that reflect the filter's effectiveness for its intended purpose, and to yield repeatable (i.e. controlled and statistically valid) results.

Figure 1 characterizes an email filter's actual usage. Incoming email messages are received by the filter, which puts them into one of two files - the ham[1] file (*in box*) and the spam file (*quarantine*). The user regularly reads the ham file, rejects any spam messages (which have been misfiled by the filter), and reads or otherwise deals with the remaining ham messages. The human may also report the misfiled spam to the filter. Occasionally (perhaps rarely or never) the spam file is searched for ham messages that have been misfiled. The human may also report such ham misfilings to the filter. The filter may use this feedback, as well as external resources such as blacklists, to improve its effectiveness.

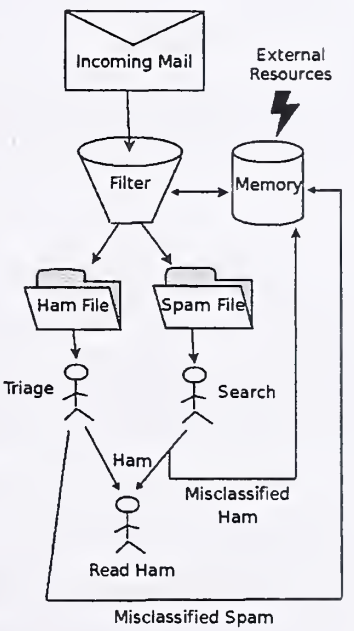

Figure 1: Real Filter Usage

The filter's effectiveness for its intended purpose has two principal aspects: the extent to which ham is placed in the ham file (not the spam file) and the extent to which spam is placed in the spam file (not the ham file). It is convenient to quantify the filter's failures in these two aspects: the *ham misclassification percentage* (*hm%*) is the fraction of all ham delivered to the spam file; the *spam misclassification percentage* (*sm%*) is the fraction of all spam delivered to the ham file. A filter is effective to the extent that it minimizes both ham and spam misclassification; however, the two have disparate impacts on the user. Spam misclassification reflects directly the extent to which the filter falls short of its intended purpose – to detect spam. Spam misclassification inconveniences and annoys the user, and may, by cluttering the ham file, cause the user to overlook important messages. Ham misclassification, on the other hand, is an undesirable side-effect of spam filtering. Ham misclassification inconveniences the user and risks loss of important messages. This risk is difficult to quantify as it depends on (1) how likely the user is to notice a ham misclassification, and (2) the importance to the user of the misclassified ham. In general, ham

[1] *Ham* denotes non-spam. Spam is defined to be "Unsolicited, unwanted email that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the recipient."

[2] An analogy may be drawn with automobile safety and fuel efficiency standards. *Deaths per 100 million km* and *litres per 100 km* are used to measure these aspects of automobile design. It is desirable to minimize both, but dimensionally meaningless to sum them or to combine them by some other linear formula.

misclassification is considerably more deleterious than spam misclassification. Because they measure qualitatively different aspects of spam filtering[2], the spam track avoids quantifying the relative importance of ham and spam misclassification.

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 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 ($(1 - ROCA)\%$).

For the reasons stated above, accuracy (percentage of correctly classified mail, whether ham or spam) is inconsistent with the effectiveness of a filter for its intended purpose[3], and is not reported here. A single quality measure, based only on the filter's binary ham/spam classifications, is nonetheless a desirable objective. To this end, spam track 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[4]. 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.

In addition to $(1 - ROCA)\%$ and *lam%*, which are threshold-neutral, the appendix reports *sm%* for various values of *hm%*, and *hm%* for various values of *sm%*. One of these statistics – *sm%* at $hamm\% = 0.1$ (denoted $h = .1$) – was chosen as indicative of overall filter effectiveness and included in comparative summary tables.

It may be argued that the filter's behaviour and the user's expectation evolve during filter use. A filter's classification performance may improve (or degrade) with use. A user may be more tolerant of errors that are made early in the filter's deployment. The spam track includes two approaches to measuring the filter's learning curve: (1) piecewise approximation and logistic regression are used to model *hm%* and *sm%* as a function of the number of messages processed; (2) cumulative *(1-ROCA)%* is given as a function of the number of messages processed.

In support of repeatability, the incoming email sequence and gold standard adjudications are fixed before filter testing. External resources are not available to the filters[5] during testing. For each measure and each corpus, 95% confidence limits are computed based on the assumption that the corpus was randomly selected from some source population with the same characteristics. *hm%* and *sm%* limits are computed using exact binomial probabilities. *lam%* limits are computed using logistic regression. *(1-ROCA)%* limits are computed using 100 bootstrap samples to estimate the standard error of $(1 - ROCA)\%$.

2 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.

TREC 2005 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.

[3] Optimizing accuracy incents filters to use threshold values that are clearly at odds with the their intended purpose.[3]

[4] 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] Nevertheless, participants are at liberty to embed an unbounded quantity of prior data in their filter submissions. Within the framework it would be possible to capture and include blacklists, DNS servers, known-spam signatures, and so on, thus simulating many external resources.

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 constraints were not rigidly enforced and, in the case of run-time, exceeded by orders of magnitude by some filters. Track guidelines indicated that the largest email sequence would not exceed 100,000 messages. This limit was exceeded as well – the largest consisted of 172,000 messages – but all filters appeared to be able to handle this size, given sufficient time. All but two participant filters – tamSPAM3 and tamSPAM4, which took 22 days and 12 days respectively to process the 49,000-message Mr. X corpus – were run on all corpora.

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 communicates the gold standard judgement to the filter before proceeding to the next message.

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.

3 Test Corpora

It is a simple matter to capture all the email delivered to a recipient or a set of recipients. Using this captured email in a public corpus, as for the other TREC tasks, is not so simple. Few individuals are willing to publish their email, because doing so would compromise their privacy and the privacy of their correspondents. A choice must be made between using a somewhat artificial public collection of messages and using a more realistic collection that must be kept private. The 2005 spam track explores this tradeoff by using both public and private collections. Participants ran their filters on the public data and submitted their results, in accordance with TREC tradition. In addition, participants submitted their filter implementations, which were run on private data by the proprietors of the data.

To form a test corpus, captured email must be augmented with gold-standard judgements. The track's definition of spam is *"Unsolicited, unwanted email that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the recipient."* The gold standard represents, as accurately as is practicable, the result of applying this definition to each message in the collection. The gold standard plays two distinct roles in the testing framework. One role is as a basis for evaluation. The gold standard is assumed to be *truth* and the filter is deemed correct when it agrees with the gold standard. The second role is as a source of user feedback. The toolkit communicates the gold standard to the filter for each message after the filter has been run on that message.

Human adjudication is a necessary component of gold standard creation. Exhaustive adjudication is tedious and error-prone; therefore we use a bootstrap method to improve both efficiency and accuracy. The bootstrap method begins with an initial gold standard G_0. One or more filters is run, using the toolkit and G_0 for feedback. The evaluation component reports all messages for which the filter and G_0 disagree. Each such message is re-adjudicated by the human and, where G_0 is found to be wrong, it is corrected. The result of all corrections is a new standard G_1. This process is repeated, using different filters, to form G_2, and so on, to G_n.

One way to construct G_0 is to have the recipient, in the ordinary course of reading his or her email, flag spam; unflagged email would be assumed to be ham. Or the recipient could use a spam filter and flag the spam filter's errors; unflagged messages would be assumed to be correctly classified by the filter. Where it is not possible to capture judgements in real time – as for all public collections to which we have access – it is necessary to construct G_0 without help from the recipient. This can be done by training a filter on a subset of the messages (or by using a filter that requires no training) and running the filter with no feedback.

3.1 Public Corpus – trec05p-1

Public Corpora

	Ham	Spam	Total
trec05p-1/full	39399	52790	92189
trec05p-1/ham25	9751	52790	62541
trec05p-1/ham50	19586	52790	72376
trec05p-1/spam25	39399	13179	52578
trec05p-1/spam50	39399	26283	65682

Private Corpora

	Ham	Spam	Total
Mr X	9038	40048	49086
S B	6231	775	7006
T M	150685	19516	170201
Total	165954	60339	226293

Table 1: Corpus Statistics

In the course of the Federal Energy Regulatory Commission's investigation, more than 1 million messages and files from the email folders of 150 Enron employees were released to the public. A digest of these files[6] was investigated as an email collection, but proved unsuitable as a large number of files did not appear to be email messages; those that were had been reformatted, deleting headers, markup, and attachments, and replacing original *message-ids* with synthetic ones. The files used in the collection were fetched directly from FERC [4]. Of these files, some 100,000 were email messages with headers; however, only 43,000 had had a "Received:" line indicating that the headers were (more-or-less) complete. These 43,000 messages form the core of the trec05p-1 public corpus.

G_0 was constructed using Spamassassin 2.63 with user feedback disabled. Subsequent iterations used a number of filters – Spamassassin, Bogofilter, Spamprobe and crm114, interleaved with human assessments for all cases in which the filter disagreed with the current gold standard. This process identified about 5% of the messages as spam.

It was problematic to adjudicate many messages because it was difficult to glean the relationship between the sender and the receiver. In particular, the collection has a preponderance of sports betting pool announcements, stock market tips, and religious bulk mail that was initially adjudicated as spam but later re-adjudicated as ham. Advertising from vendors whose relationship with the recipient was tenuous presented an adjudication challenge.

During this process, the need arose to view the messages by sender; for example, once the adjudicator decides that a particular sports pool is indeed by subscription, it is more efficient and probably more accurate to adjudicate all messages from the same sender at one time. Similarly, in determining whether or not a particular "newsletter" is spam, it is desirable to identify all of its recipients. This observation occasioned the design and use of a new tool for adjudication – one that allows the adjudicator to use full-text retrieval to look for evidence and to ensure consistent judgements.

The 43,000 Enron messages were augmented by approximately 50,000 spam messages collected in 2005. The headers of these messages were altered so as to appear that they were delivered to the Enron mail server during the same time frame (summer 2001 through summer 2002). "To:" and "From:" headers, as well as the message bodies, were altered to substitute the names and email addresses of Enron employees for those of the original recipients. Spamassassin 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.

A final iteration of bootstrap process was effected to produce the final gold standard.

In addition to the full public corpus, four subsets were defined. These subsets use the same email collection and gold standard judgements, but include only a subset of the index entries so as to reflect different proportions of ham and spam. *trec05p-1/spam50* contains all of the ham and 50% of the spam from the full corpus; *trec05p-1/spam25* contains all of the ham and 25% of the spam. Similarly *trec05p-1/ham50* contains all of the spam and 50% of the ham, while *trec05p-1/ham25* contains all of the spam and 25% of the ham. All subsets were chosen at random. The numbers of ham and spam in each corpus are reported in table 1.

3.2 Private Corpus – Mr. X

The Mr. X corpus was created by Cormack and Lynam in 2004[3]. The email collection consists of the 49086 messages received by an individual, X, from August 2003 through March 2004. X has had the same email address for twenty years; variants of X's email address appear on the Web and in Usenet archives. X has subscribed to services and purchased goods on the Internet. X used a spam filter – Spamassassin 2.60 – during the period in question, and reported observed misclassifications to the filter. G_0 was captured from the filter's database. Table 2 illustrates the five revision steps forming G_1 through G_5, the final gold standard. $S \rightarrow H$ is the number of message classifications revised from spam to ham; $H \rightarrow S$ is the opposite. Note that G_0 had 421 spam messages incorrectly classified as ham. Left uncorrected, these errors would cause the evaluation kit to over-report the false positive rate of the filters by this a mount – more than an order of magnitude for the best filters. In other words, the results captured from user feedback alone – G_0 – were not accurate enough to form a useful gold standard. G_5, on the other hand, appears to be sufficiently accurate; systematic inspection of the 2004 results and of the 2005 spam track results reveals no gold standard errors – any that may persist do not contribute materially to the results.

3.3 Private Corpus – S. B.

The S. B. corpus consists of 7,006 messages (89% ham, 11% spam) received by an individual in 2005. The majority of all ham messages stems from 4 mailing lists (23%, 10%, 9%, and 6% of all ham messages) and private messages received from 3 frequent correspondents (7%, 3%, and 2%, respectively), while the vast majority of the spam messages (80%) are traditional spam: viruses, phishing, pornography, and Viagra ads.

	$S \rightarrow H$	$H \rightarrow S$
$G_0 \rightarrow G_1$	0	278
$G_1 \rightarrow G_2$	4	83
$G_2 \rightarrow G_3$	0	56
$G_3 \rightarrow G_4$	10	15
$G_4 \rightarrow G_5$	0	0
$G_0 \rightarrow G_5$	8	421
G_5	$\|H\| = 9038$	$\|S\| = 40048$

Table 2: Mr. X Bootstrap Gold Standard Iterations

Starting from a manual preclassification of all emails, performed when each message arrived in the mailbox, the gold standard was created by running at least one spam filter from each participating group and manually reclassifying all messages for which at least one of the filters disagreed with the preclassification. During this process, 95% of all spam messages and 15% of all ham messages were manually re-adjudicated, and reclassified as necessary. Genre classification was done using a mixture of email header pattern matching (for mailing lists and newsletters) and manual classification.

3.4 Private Corpus – T. M.

The T. M. corpus [7] includes personal email, from all accounts owned by an individual, including all mail received (except for spam filtered out by gmail to the gmail address). There are 170,201 messages in total. Messages were manually classified as they arrived, and the classifications were verified them by running his filter over the corpus and manually examining all false positives, false negatives and unsures until there were no more errors. Further verification was effected by running Bogofilter, SpamProbe, SpamBayes and CRM114 (in the TREC setup) over the corpus, manually examining all false positives and false negatives. The corpus ranges from Tue, 30 Apr 2002 to Wed, 6 Apr 2005.

4 Spam Track Participation

Group	Filter Prefixes
Beijing University of Posts and Telecommunications	kidSPAM1, kidSPAM2, kidSPAM3, kidSPAM4
Chinese Academy of Sciences (ICT)	ICTSPAM1, ICTSPAM2, ICTSPAM3, ICTSPAM4
Dalhousie University	dalSPAM1, dalSPAM2, dalSPAM3, dalSPAM4
IBM Research (Segal)	621SPAM1, 621SPAM2, 621SPAM3
Indiana University	indSPAM1, indSPAM2, indSPAM3, indSPAM4
Jozef Stefan Institute	ijsSPAM1, ijsSPAM2, ijsSPAM3, ijsSPAM4
Laird Breyer	lbSPAM1, lbSPAM2, lbSPAM3, lbSPAM4
Tony Meyer (Massey University in appendix)	tamSPAM1, tamSPAM2, tamSPAM3, tamSPAM4
Mitsubishi Electric Research Labs (CRM114)	crmSPAM1, crmSPAM2, crmSPAM3, crmSPAM4
Pontificia Universidade Catolica Do Rio Grande Do Sul	pucSPAM1, pucSPAM2, pucSPAM3
Universite Paris-Sud	azeSPAM1, azeSPAM2
York University	yorSPAM1, yorSPAM2, yorSPAM3, yorSPAM4

Table 3: Participant filters

The filter evaluation toolkit was made available in advance to participating groups. In addition to the testing and evaluation components detailed above, the toolkit included a sample public corpus, derived from the Spamassassin Corpus [10], and eight open-source sample filter implementations: Bogofilter [9], CRM114 [12], DSPAM [13], dbacl [1], Popfile [5], Spamassassin [11], SpamBayes [8], and Spamprobe [2].

Participating groups were required to configure their filters to conform to the toolkit, and to submit a pilot implementation which was run by the track coordinators on the supplied corpus and also on a 150-message sample of Enron email. Thirteen groups submitted pilot filters; results and problems with the pilot runs were reported back to these groups.

Each group was invited to submit up to four filter implementations for final evaluation; twelve groups submitted a total of 44 filters for final evaluation. Groups were asked to prioritize their submissions in case insufficient resources were available

Filter	*Run Prefix*	*Configuration*
Bogofilter	bogofilter	0.92.2
DSPAM	dspam-tum	3.4.9, train-until-mature
	dspam-toe	3.4.9, train-on-errors
	dspam-teft	3.4.9, train-on-everything
Popfile	popfile	0.22.2
Spamassassin	spamasasb	3.0.2, Bayes component only
	spamasasv	3.0.2, Vanilla (out of the box)
	spamasasx	3.0.2, Mr. X configuration
Spamprobe	spamprobe	1.0a

Table 4: Non-participant filters

to test all filters on all corpora, but it was not necessary to use this information – all but two of the 44 filters, mentioned above, were run on all private corpora.

Following the filter submissions, the public corpus trec05p-1 was made available to participants, who were required to run their filters, as submitted, on trec05p-1/full and submit the results. Participants were also encouraged to run their filters on the subset corpora.

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 3 shows the identifier prefix for each submitted filter.

4.1 Non-participant Runs

For comparison, revised versions of the open-source filters supplied with the toolkit were run on the spam track corpora. The authors of three – crm114, dbacl, and Spambayes – were spam track participants. The authors of the remaining five – Bogofilter, DSPAM, Popfile, Spamassassin, and Spamprobe were approached to suggest revisions or variants of their filters. These versions were tested in the same manner as the participant runs. Table 4 illustrates each non-participant filter.

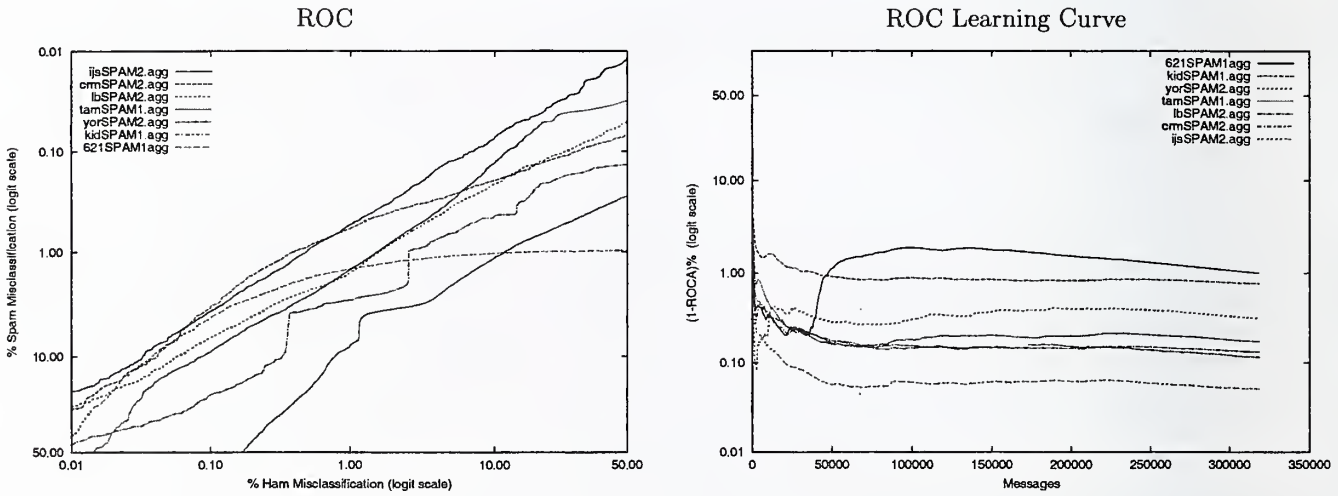

Figure 2: Aggregate

4.2 Aggregate Runs

The subject filters were run separately on the various corpora. That is, each filter was subject to (up to) eight test runs. The four full corpora – trec05p-1/full, mrx, sb, and tm – provide the primary results for comparison. For each filter, and *aggregate run* was created combining its results on the four corpora as if they were one. The evaluation component of the

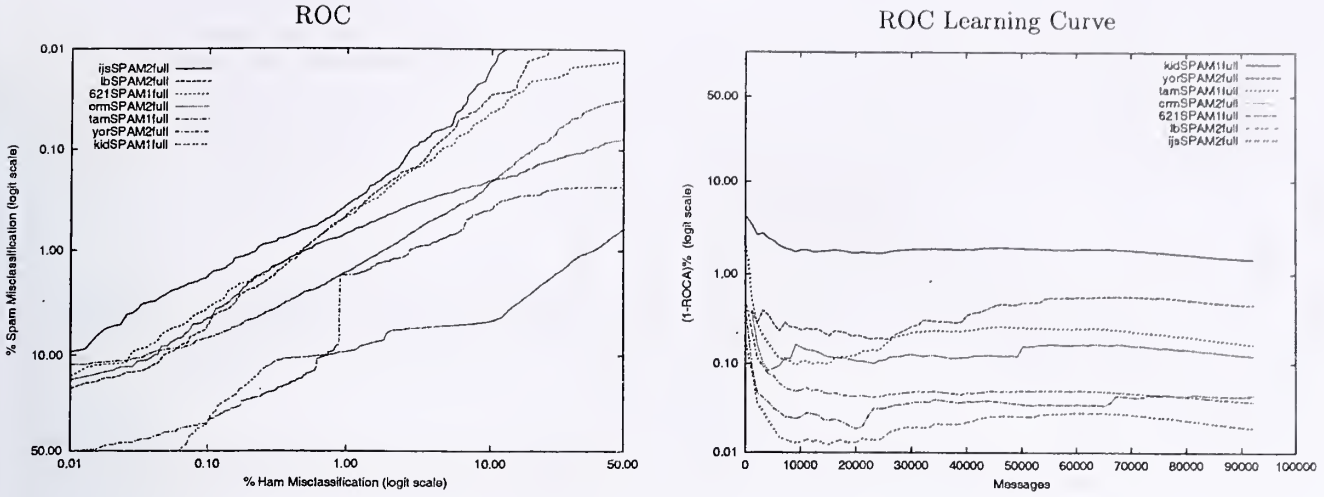

Figure 3: trec05–1/full

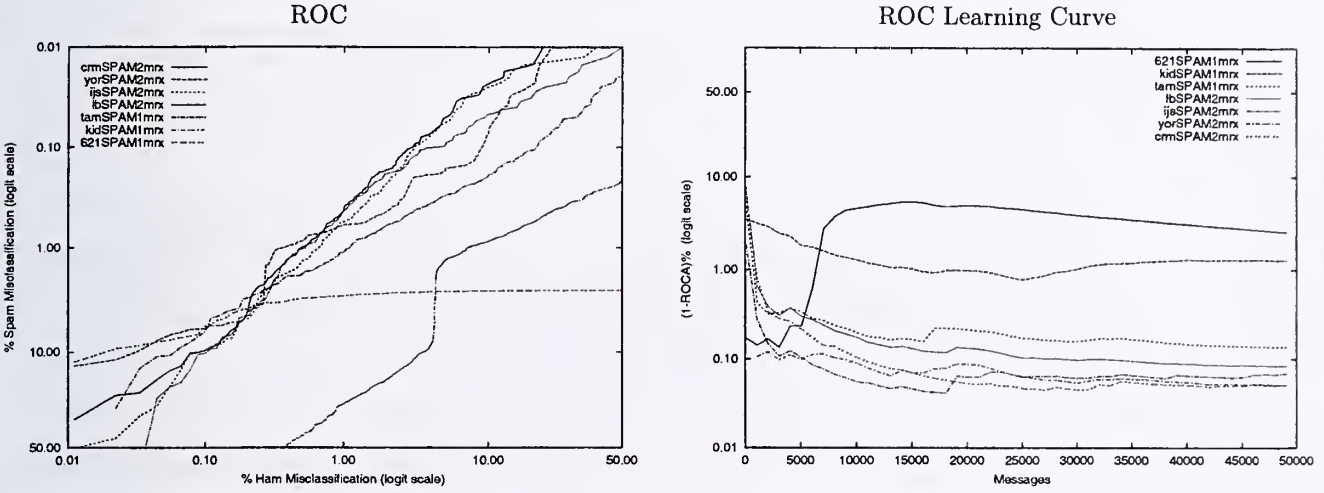

Figure 4: Mr X

toolkit was run on the aggregate results, consisting of 318,482 messages in total – 113,129 spam and 205,253 ham. The summary results on the aggregate runs provide a composite view of the performance on all corpora.

5 Results

Table 5 presents the three measures of the binary classification measures: *hm%, sm%,* and *lam%.* Table 6 presents three summary measurements of filter quality – *(1-ROCA)%, h=.1%, and lam%.* Table 7 shows the relative ranks achieved by the filters according to each of the fifteen summary measures. The tables show each filter's performance on each of the four full corpora, and in the aggregate, ordered by aggregate *(1-ROCA)%.* More detailed results for each run, including confidence limits and graphs, may be found in the notebook appendix.

Figure 2 shows the ROC curves for the best seven participant runs ranked by (1-ROCA)%, and restricted to one run (the best) per participant. ijsSPAM2 dominates the other curves over most regions. However, if one considers the intercept with the 0.10% ham misclassification line, crmSPAM2 is slightly (but not significantly) higher. This difference is reflected in the different rankings shown in table 7. It may be argued that this intercept accurately reflects the usefulness of the filter for its intended purpose. On the other hand, a broad ROC curve may be argued to reflect good filtering performance. Indeed, the crm group indicated that the falloff of the curve was due to a bug they discovered in the course of their TREC

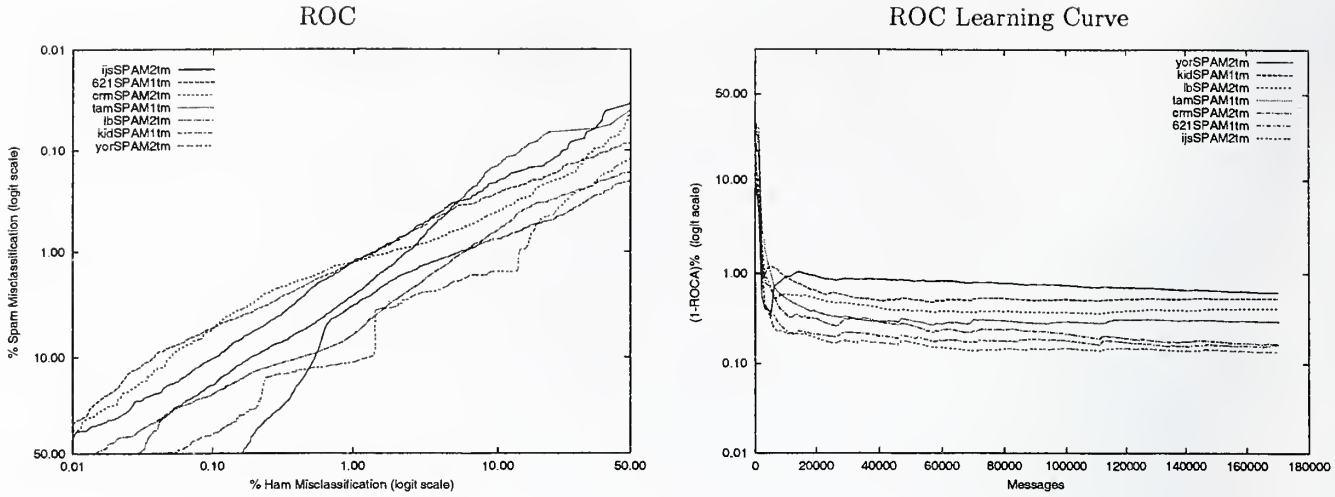

Figure 5: T. M.

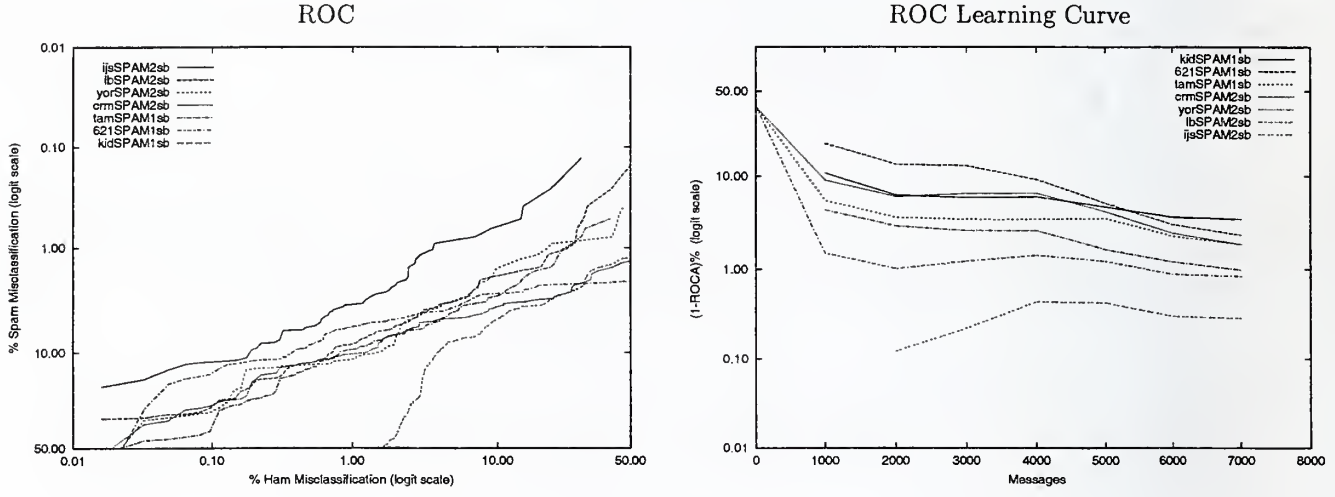

Figure 6: S. B.

participation. 621SPAM1 demonstrates a severe falloff, also due to a bug – this filter failed on every message larger than 100KB. Figures 3 through 6 show the curves for the same filters on the four primary corpora. Figure 7 shows the ROC curves for the non-participant aggregate runs; additionally, for comparison, the best participant run.

Learning curves for the aggregate and four major corpora are also shown in figures 2 through 6. These curves show (1-ROCA)% as a function of the number of messages classified. The curves appear to indicate that the filters have reached steady-state performance. Instantaneous ham and spam learning curves for each run are given in the notebook appendix.

Table 11 gives a *genre* classification for each misclassified message in the S. B. Corpus. Genre classification may be useful to assess the impact of misclassification; for instance, a misclassified personal message or a message from a frequent correspondent is more likely to have serious negative consequences than a misclassified newsletter article. In addition, genre classification may give insight into the nature of messages that are difficult to classify. The ham genres are:

- *Automated.* Sent by software to the recipient, perhaps as part of an Internet transaction.
- *Commercial.* Commercial email not considered spam.
- *Encrypted.* Personal or other sensitive email, sent in an encrypted format.
- *Frequent.* Email from a frequent correspondent.

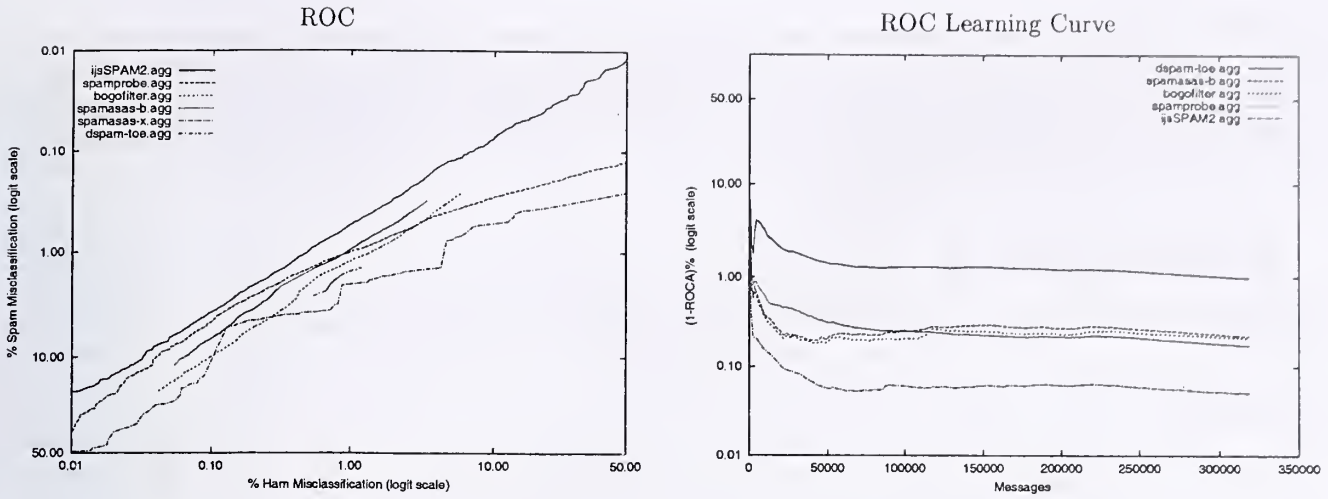

Figure 7: Non-participant Aggregate

- *List.* Email from a mailing list
- *Newsletter.* Message from a subscribed-to news service.
- *Personal.* Personal individual correspondence.

The spam genres are:

- *Automated.* Unwelcome messages sent automatically to the recipient.
- *List.* Spam delivered via a mailing list to which the recipient is subscribed.
- *Newsletter.* An unwelcome newsletter to which the recipient did not subscribe.
- *Phishing.* Fraudulent email misrepresenting its origin or purpose.
- *Sex.* Pornography or other sexually-related spam.
- *Virus.* An email message containing a virus.

6 Conclusions

Notwithstanding a few operational issues which occasioned extensions to deadlines, relaxation of limits, and patches to filters, the submission mechanism worked satisfactorily. Participants submitted filters to the track, and also ran the same filters on public data received by the track. The public corpus appears to have yielded comparable results to those achieved on the private corpora – preliminary analysis shows that the statistical differences between the results on private and public corpora appear to be no larger than those among the private corpora. This observation contradicts the authors' prior prediction, which was that large anomalies would be apparent in the public corpus results. Further post-hoc analysis will likely uncover some artifacts of the public corpus that worked either to the filters' advantage or disadvantage.

The results presented here indicate that content-based spam filters can be quite effective, but not a panacea. Misclassification rates are easily observable, even with the smallest corpus of about 8,000 messages. The results call into question a number of public claims both as to the effectiveness and ineffectiveness of "Bayesian" and "statistical" spam filters.

The filters did not, in general, appear to be seriously disadvantaged by the lack of an explicit training set. Their error rates converged quickly, and the overall misclassification percentages were not dominated by early errors. In any event, the use of a training set would have been inconsistent with the track objective of modelling real usage as closely as possible.

TREC 2005 did not afford the filters on-line access to external resources, such as black lists, name servers, and the like. Participants could have included, but did not, archived versions of these resources with their submissions. No aspect of

the toolkit or evaluation measures precludes the use of on-line resources; privacy and repeatability considerations excluded them at TREC. The efficacy of these resources remains an open question, notwithstanding public claims in this regard.

The public corpus will be made generally available, subject to a standard TREC usage agreement that proscribes disclosure of information that would compromise its utility as a test corpus. It may be desirable, before the corpus is made generally available, to use it in another round of blind testing.

7 Acknowledgements

The authors thank Tony Meyer and Stefan Buettcher for their invaluable contributions to this effort.

References

[1] Breyer, L. Laird breyer's free software - dbacl. http://www.lbreyer.com/gpl.html, 2005.

[2] Burton, B. Spamprobe - a fast bayesian spam filter. http://spamprobe.sourceforge.net, 2002.

[3] Cormack, G., and Lynam, T. A study of supervised spam detection applied to eight months of personal email. http://http://plg.uwaterloo.ca/ gvcormac/spamcormack.html, 2004.

[4] FERC. Information released in enron investigation. http://fercic.aspensys.com/members/manager.asp, 2003.

[5] Graham-Cumming, J. Popfile. http://popfile.sourceforge.net/, 2004.

[6] Klimt, B., and Yang, Y. Introducing the enron corpus. In *First Conference on Email and Anti-Spam (CEAS)* (2004).

[7] Meyer, T. Email classification. http://www.massey.ac.nz/ tameyer/research/spambayes/index.html, 2005.

[8] Peters, T. Spambayes: Bayesian anti-spam classifier in python. http://spambayes.sourceforge.net/, 2004.

[9] Raymond, E. S., Relson, D., Andree, M., and Louis, G. Bogofilter. http://bogofilter.sourceforge.net/, 2004.

[10] spamassassin.org. The spamassassin public mail corpus. http://spamassassin.apache.org/publiccorpus, 2003.

[11] spamassassin.org. Welcome to spamassassin. http://spamassassin.apache.org, 2005.

[12] Yerazunis, W. S. CRM114 - the controllable regex mutilator. http://crm114.sourceforge.net/, 2004.

[13] Zdziarski, J. A. The DSPAM project. http://www.nuclearelephant.com/projects/dspam/, 2004.

	Aggregate			trec05p-1/full			Mr. X			S. B.			T. M.		
Filters	hm%	sm%	lam%	hm%	sm%	lam%	hm%	sm%	lam%	hm%	sm%	lam%	hm%	sm%	lam%
ijsSPAM2	0.38	1.24	0.69	0.23	0.95	0.47	1.52	0.34	0.72	0.16	11.74	1.44	0.36	3.43	1.12
ijsSPAM1	0.39	1.22	0.69	0.25	0.93	0.48	1.54	0.39	0.77	0.35	11.35	2.09	0.36	3.31	1.10
ijsSPAM4	0.46	1.28	0.77	0.37	0.91	0.58	1.44	0.54	0.88	0.30	12.77	2.07	0.43	3.34	1.21
ijsSPAM3	0.64	1.32	0.92	0.26	0.97	0.51	1.33	0.33	0.66	0.59	11.10	2.66	0.70	3.87	1.66
crmSPAM2	0.35	1.08	0.62	0.62	0.87	0.73	1.50	0.24	0.60	0.30	13.55	2.14	0.21	2.91	0.79
crmSPAM3	0.73	1.40	1.01	2.56	0.15	0.63	0.58	1.66	0.98	0.34	7.61	1.64	0.28	3.99	1.08
crmSPAM4	0.37	1.05	0.62	0.91	0.25	0.47	0.67	0.91	0.79	0.39	6.97	1.67	0.21	3.26	0.83
lbSPAM2	0.38	2.75	1.03	0.51	0.93	0.69	1.63	0.23	0.62	0.03	33.16	1.25	0.29	11.63	1.90
lbSPAM1	0.34	2.46	0.91	0.41	0.90	0.61	1.14	0.28	0.57	0.05	36.13	1.62	0.28	9.80	1.72
tamSPAM1	0.25	4.43	1.07	0.26	4.10	1.05	0.28	2.55	0.84	0.14	27.48	2.29	0.25	8.25	1.49
spamprobe	0.14	3.35	0.70	0.15	2.11	0.57	0.41	0.85	0.59	0.05	42.06	1.84	0.13	10.31	1.21
tamSPAM2	0.45	2.63	1.10	0.85	1.45	1.11	1.43	1.75	1.58	1.04	9.29	3.18	0.27	7.36	1.44
bogofilter	0.09	10.86	1.02	0.01	10.47	0.30	0.08	6.51	0.73	0.00	73.03	1.15	0.11	18.41	1.57
spamasas-b	0.31	2.17	0.83	0.25	1.29	0.57	0.49	1.00	0.70	0.06	25.68	1.47	0.33	6.00	1.43
lbSPAM3	0.68	2.81	1.39	0.83	1.05	0.94	6.84	0.35	1.57	0.47	42.45	5.55	0.29	11.04	1.85
crmSPAM1	0.80	3.75	1.74	1.84	1.65	1.74	4.22	0.50	1.46	0.37	13.42	2.34	0.33	15.72	2.44
lbSPAM4	0.59	4.66	1.67	0.91	3.87	1.89	4.86	1.21	2.44	0.26	49.94	4.82	0.26	12.06	1.87
yorSPAM2	0.38	3.91	1.23	0.92	1.74	1.27	0.34	1.03	0.60	0.14	23.64	2.07	0.25	14.90	2.05
spamasas-x	0.13	5.39	0.85	0.15	3.16	0.70	0.14	2.28	0.58	0.00	14.84	0.29	0.13	17.43	1.61
kidSPAM1	0.93	8.60	2.88	0.91	9.40	2.99	4.02	9.10	6.08	3.37	13.57	6.89	0.65	5.24	1.86
dspam-toe	0.58	1.88	1.05	1.04	0.99	1.01	1.94	0.59	1.07	0.05	30.97	1.45	0.40	5.78	1.55
621SPAM1	2.20	1.23	1.65	2.38	0.20	0.69	2.31	2.77	2.53	1.57	5.29	2.90	2.17	0.74	1.27
621SPAM3	0.70	12.58	3.08	3.14	0.17	0.73	1.73	2.87	2.23	0.56	7.48	2.09	0.00	66.27	0.95
yorSPAM4	1.29	2.98	1.96	2.99	1.36	2.02	5.20	0.45	1.55	0.77	91.35	22.26	0.63	9.04	2.45
dspam-tum	0.31	2.57	0.89	0.26	1.79	0.69	1.81	0.57	1.02	0.05	35.23	1.59	0.24	7.48	1.37
dspam-teft	0.26	2.93	0.87	0.26	1.79	0.69	1.85	0.53	0.99	0.00	44.26	0.63	0.17	9.32	1.31
yorSPAM3	1.16	2.29	1.63	1.29	1.20	1.25	4.41	0.65	1.71	1.36	15.32	4.76	0.92	8.11	2.78
dalSPAM3	5.44	8.65	6.87	6.80	6.23	6.51	3.44	9.79	5.86	4.03	13.29	7.43	5.27	12.63	8.23
yorSPAM1	1.32	2.85	1.94	2.44	2.43	2.44	4.96	0.55	1.67	1.19	13.81	4.20	0.82	8.28	2.65
dalSPAM1	0.92	18.93	4.44	1.17	21.07	5.33	1.17	13.83	4.18	1.35	38.19	8.42	0.82	22.82	4.70
dalSPAM2	5.40	9.64	7.24	5.34	7.52	6.34	3.12	11.33	6.03	4.73	12.39	7.73	5.58	11.83	8.17
kidSPAM4	2.94	5.05	3.86	9.74	6.57	8.01	5.31	2.39	3.57	5.75	18.09	10.40	0.91	5.88	2.34
kidSPAM3	0.75	11.11	2.99	0.82	12.49	3.33	3.03	10.27	5.64	2.86	24.42	8.89	0.51	8.58	2.15
kidSPAM2	0.84	9.71	2.92	0.87	10.53	3.11	2.71	9.89	5.24	3.40	16.15	7.62	0.61	6.85	2.08
ICTSPAM2	4.31	9.80	6.54	8.33	8.03	8.18	4.51	3.42	3.93	1.08	15.74	4.31	3.38	27.41	10.31
dalSPAM4	2.92	11.66	5.93	2.69	4.50	3.49	2.18	14.40	5.77	1.89	40.90	10.36	3.07	24.23	9.14
indSPAM3	2.49	8.74	4.71	1.09	7.66	2.93	1.81	4.62	2.90	1.83	37.03	9.48	2.92	18.98	7.74
pucSPAM0	2.21	8.06	4.27	3.41	5.10	4.18	4.93	2.26	3.35	1.44	24.13	6.39	1.77	27.34	7.61
indSPAM1	2.54	13.57	6.01	0.82	15.16	3.70	1.47	5.83	2.95	1.83	41.03	10.22	3.08	24.05	9.11
pucSPAM1	2.30	8.38	4.44	3.57	5.33	4.36	5.97	2.72	4.05	1.03	17.29	4.45	1.80	27.87	7.77
621SPAM2	14.59	4.50	8.23	55.06	1.07	10.32	25.82	2.87	9.21	2.47	6.84	4.13	3.83	17.02	8.29
pucSPAM2	2.58	7.17	4.33	3.35	5.00	4.10	6.07	2.77	4.12	2.47	40.90	11.70	2.17	20.73	7.08
ICTSPAM1	23.16	15.20	18.86	5.69	20.85	11.19	4.47	2.37	3.26	1.01	18.06	4.53	29.76	26.12	27.91
ICTSPAM3	13.11	27.33	19.24	14.10	28.22	20.26	9.57	3.91	6.16	6.61	19.35	11.53	13.33	73.29	39.38
ICTSPAM4	62.14	16.02	35.88	8.18	24.89	14.66	6.68	4.10	5.24	6.31	14.97	9.82	81.88	16.54	48.62
azeSPAM1	30.78	4.21	12.26	64.84	4.57	22.92	47.81	2.28	12.76	57.90	9.16	27.14	19.73	6.97	11.95
spamasas-v	-	-	-	0.06	39.51	1.87	0.02	11.70	0.54	0.02	72.13	2.00	-	-	-
popfile	-	-	-	0.92	1.26	0.94	0.96	0.49	0.69	0.14	22.97	2.03	-	-	-
tamSPAM4	-	-	-	-	-	-	0.96	0.89	0.92	3.92	6.19	4.93	-	-	-
tamSPAM3	-	-	-	0.22	4.46	1.01	0.82	1.85	1.23	6.29	3.64	4.79	-	-	-
indSPAM4	-	-	-	-	-	-	1.28	7.49	3.14	0.93	35.23	6.67	0.34	16.74	2.54
indSPAM2	-	-	-	-	-	-	2.66	3.09	2.86	0.03	100.00	99.99	2.87	21.41	8.24
azeSPAM2	-	-	-	-	-	-	8.54	25.35	15.12	8.04	59.48	26.38	0.63	36.84	5.75

Table 5: Misclassification Summary

	Aggregate			trec05p-1/full			Mr. X			S. B.			T. M.		
Filters	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%
ijsSPAM2	0.051	3.78	0.69	0.019	1.78	0.47	0.069	9.72	0.72	0.285	12.13	1.44	0.135	10.31	1.12
ijsSPAM1	0.054	3.73	0.69	0.021	1.84	0.48	0.069	13.63	0.77	0.372	15.87	2.09	0.155	9.88	1.10
ijsSPAM4	0.058	4.91	0.77	0.025	2.22	0.58	0.063	8.68	0.88	0.422	17.03	2.07	0.167	12.66	1.21
ijsSPAM3	0.064	5.54	0.92	0.022	1.84	0.51	0.050	3.56	0.66	0.475	21.29	2.66	0.181	14.49	1.66
crmSPAM2	0.115	3.46	0.62	0.122	4.52	0.73	0.051	9.65	0.60	1.888	27.48	2.14	0.166	5.64	0.79
crmSPAM3	0.116	10.50	1.01	0.042	2.63	0.63	0.177	40.82	0.98	0.231	11.23	1.64	0.195	13.06	1.08
crmSPAM4	0.128	5.90	0.62	0.049	1.96	0.47	0.218	82.36	0.79	0.393	15.23	1.67	0.272	8.59	0.83
lbSPAM2	0.132	6.75	1.03	0.037	5.19	0.69	0.083	10.24	0.62	0.835	28.52	1.25	0.411	20.53	1.90
lbSPAM1	0.136	6.19	0.91	0.039	4.56	0.61	0.103	20.67	0.57	0.778	31.61	1.62	0.443	17.94	1.72
tamSPAM1	0.172	9.10	1.07	0.164	6.92	1.05	0.138	6.51	0.84	1.892	40.52	2.29	0.294	17.40	1.49
spamprobe	0.173	4.71	0.70	0.059	2.77	0.57	0.097	15.54	0.59	2.039	28.77	1.84	0.445	12.02	1.21
tamSPAM2	0.209	15.50	1.10	0.178	27.38	1.11	0.349	78.36	1.58	1.127	66.06	3.18	0.416	19.41	1.44
bogofilter	0.210	9.86	1.02	0.048	3.41	0.30	0.045	3.90	0.73	1.426	30.97	1.15	0.792	19.86	1.57
spamasas-b	0.220	6.72	0.83	0.059	2.56	0.57	0.097	6.19	0.70	1.620	19.87	1.47	0.736	15.58	1.43
lbSPAM3	0.262	29.95	1.39	0.122	22.38	0.94	0.875	95.73	1.57	2.727	98.32	5.55	0.456	22.38	1.85
crmSPAM1	0.263	12.79	1.74	0.169	10.53	1.74	0.311	81.61	1.46	2.393	23.48	2.34	0.790	23.12	2.44
lbSPAM4	0.302	17.23	1.67	0.238	22.94	1.89	0.492	58.36	2.44	1.988	52.65	4.82	0.588	19.67	1.87
yorSPAM2	0.316	21.14	1.23	0.457	34.21	1.27	0.051	6.08	0.60	0.983	30.52	2.07	0.619	39.19	2.05
spamasas-x	0.380	11.17	0.85	0.345	16.59	0.70	0.065	2.50	0.58	0.558	10.84	0.29	1.123	29.50	1.61
kidSPAM1	0.768	66.13	2.88	1.463	34.93	2.99	1.274	83.55	6.08	3.553	99.22	6.89	0.530	62.56	1.86
dspam-toe	0.987	83.68	1.05	0.773	88.76	1.01	1.109	96.23	1.07	14.149	31.61	1.45	2.626	77.16	1.55
621SPAM1	1.008	4.36	1.65	0.044	3.63	0.69	2.616	5.71	2.53	2.389	15.48	2.90	0.161	5.42	1.27
621SPAM3	1.090	7.89	3.08	0.060	7.02	0.73	2.692	4.55	2.23	2.604	17.16	2.09	0.332	6.15	0.95
yorSPAM4	1.122	81.80	1.96	0.688	84.92	2.02	1.407	96.18	1.55	58.165	98.06	22.26	1.081	78.66	2.45
dspam-tum	1.274	51.43	0.89	0.827	47.09	0.69	0.997	95.18	1.02	19.384	40.77	1.59	3.700	37.22	1.37
dspam-teft	1.383	51.60	0.87	0.827	47.09	0.69	0.942	95.17	0.99	21.428	43.35	0.63	4.263	33.79	1.31
yorSPAM3	1.491	70.88	1.63	0.861	62.13	1.25	1.993	92.07	1.71	8.234	70.13	4.76	4.366	78.42	2.78
dalSPAM3	1.873	59.50	6.87	1.491	41.00	6.51	1.613	70.03	5.86	2.845	77.16	7.43	3.090	59.70	8.23
yorSPAM1	1.917	84.38	1.94	2.032	87.24	2.44	2.632	95.76	1.67	7.237	77.16	4.20	4.400	78.76	2.65
dalSPAM1	2.097	99.15	4.44	2.348	99.75	5.33	2.240	99.31	4.18	4.614	100.00	8.42	3.085	52.08	4.70
dalSPAM2	2.100	60.64	7.24	1.674	41.92	6.34	1.824	69.41	6.03	3.293	83.48	7.73	2.898	59.84	8.17
kidSPAM4	2.606	89.15	3.86	3.990	93.74	8.01	2.326	98.23	3.57	8.042	95.22	10.40	2.473	85.34	2.34
kidSPAM3	2.741	88.23	2.99	4.167	90.62	3.33	2.822	97.67	5.64	6.360	93.67	8.89	2.653	82.11	2.15
kidSPAM2	3.003	88.29	2.92	4.544	91.65	3.11	2.738	97.64	5.24	7.020	97.29	7.62	2.749	85.16	2.08
ICTSPAM2	3.048	60.29	6.54	2.643	79.51	8.18	0.943	37.43	3.93	3.110	99.35	4.31	8.298	86.36	10.31
dalSPAM4	3.115	79.14	5.93	1.370	76.58	3.49	4.282	96.93	5.77	9.002	100.00	10.36	6.294	58.51	9.14
indSPAM3	3.168	96.99	4.71	2.822	97.35	2.93	2.321	99.31	2.90	12.454	91.10	9.48	5.843	99.41	7.74
pucSPAM0	4.030	59.56	4.27	2.083	59.71	4.18	1.910	51.00	3.35	1.408	61.81	6.39	2.925	88.94	7.61
indSPAM1	4.302	96.06	6.01	5.346	93.19	3.70	2.471	99.10	2.95	13.507	93.16	10.22	8.382	99.44	9.11
pucSPAM1	5.746	57.60	4.44	2.185	52.58	4.36	3.081	55.92	4.05	1.585	56.52	4.45	2.712	88.48	7.77
621SPAM2	6.064	54.21	8.23	11.362	28.85	10.32	6.814	59.16	9.21	3.169	61.94	4.13	2.647	47.89	8.29
pucSPAM2	6.107	99.98	4.33	1.967	51.28	4.10	3.454	78.25	4.12	5.437	73.42	11.70	3.688	99.99	7.08
ICTSPAM1	15.115	67.60	18.86	4.659	72.26	11.19	0.748	41.24	3.26	3.023	97.55	4.53	34.208	98.13	27.91
ICTSPAM3	17.637	99.17	19.24	20.485	99.39	20.26	5.328	98.50	6.16	9.985	98.71	11.53	36.233	99.75	39.38
ICTSPAM4	33.879	99.84	35.88	10.952	98.44	14.66	4.114	97.95	5.24	6.112	97.03	9.82	42.893	99.83	48.62
azeSPAM1	34.079	99.76	12.26	28.887	99.50	22.92	34.048	99.69	12 76	44.502	99.48	27.14	39.082	99.72	11.95
spamasas-v	-	-	-	0.516	31.31	1.87	0.091	4.97	0.54	5.736	68.26	2.00	-	-	-
popfile	-	-	-	0.325	7.35	0.94	0.326	86.94	0.69	2.199	24.65	2.03	-	-	-
tamSPAM4	-	-	-	-	-	-	0.159	46.24	0.92	1.421	89.68	4.93	-	-	-
tamSPAM3	-	-	-	0.183	7.64	1.01	0.257	58.80	1.23	1.934	96.49	4.79	-	-	-
indSPAM4	-	-	-	-	-	-	1.757	97.33	3.14	9.588	100.00	6.67	3.388	96.77	2.54
indSPAM2	-	-	-	-	-	-	2.804	99.75	2.86	68.572	99.87	99.99	13.462	99.44	8.24
azeSPAM2	-	-	-	-	-	-	29.765	99.95	15.12	37.739	100.00	26.38	22.625	99.89	5.75

Table 6: Summary Results

	Aggregate			trec05p-1/full			Mr. X			S. B.			T. M.		
Filters	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%
ijsSPAM2	1	3	3	1	1	2	7	12	11	2	3	5	1	6	6
ijsSPAM1	2	2	3	2	2	4	7	14	13	3	6	17	2	5	5
ijsSPAM4	3	6	6	4	5	8	5	10	16	5	7	15	5	8	7
ijsSPAM3	4	7	12	3	2	5	2	2	8	6	10	22	6	10	18
crmSPAM2	5	1	1	14	11	16	3	11	5	17	13	19	4	2	1
crmSPAM3	6	15	13	7	7	10	16	18	18	1	2	10	7	9	4
crmSPAM4	7	8	1	10	4	2	17	31	14	4	4	11	8	4	2
lbSPAM2	8	11	15	5	13	11	9	13	7	9	14	4	11	17	23
lbSPAM1	9	9	11	6	12	9	13	16	2	8	18	9	13	13	19
tamSPAM1	10	13	17	16	14	22	14	9	15	18	20	20	9	12	14
spamprobe	11	5	5	11	8	6	11	15	4	21	15	12	14	7	7
tamSPAM2	12	18	18	18	22	23	21	29	26	11	27	24	12	14	13
bogofilter	13	14	14	9	9	1	1	3	12	14	17	3	21	16	16
spamasas-b	14	10	7	11	6	6	11	8	10	16	9	7	19	11	12
lbSPAM3	15	21	20	14	20	18	24	37	25	26	44	34	15	18	20
crmSPAM1	16	17	24	17	18	26	19	30	23	24	11	21	20	19	28
lbSPAM4	17	19	23	20	21	28	22	23	30	20	23	32	17	15	22
yorSPAM2	18	20	19	23	25	25	3	7	5	10	16	15	18	23	24
spamasas-x	19	16	8	22	19	15	6	1	3	7	1	1	23	20	17
kidSPAM1	20	30	27	31	26	32	29	32	49	32	46	37	16	29	21
dspam-toe	21	35	16	26	40	20	28	40	21	47	18	6	25	30	15
621SPAM1	22	4	22	8	10	11	40	6	31	23	5	23	3	1	9
621SPAM3	23	12	30	13	15	16	42	4	29	25	8	17	10	3	3
yorSPAM4	24	34	26	25	38	29	30	39	24	52	43	50	22	32	29
dspam-tum	25	22	10	27	29	11	27	36	20	48	21	8	36	22	11
dspam-teft	26	23	9	27	29	11	25	35	19	49	22	2	37	21	10
yorSPAM3	27	32	21	29	34	24	35	34	28	41	29	30	38	31	32
dalSPAM3	28	26	40	32	27	42	31	27	47	27	31	38	33	27	40
yorSPAM1	29	36	25	35	39	30	41	38	27	39	31	26	39	33	31
dalSPAM1	30	42	34	38	49	40	36	49	42	33	50	41	32	25	33
dalSPAM2	31	29	41	33	28	41	33	26	48	31	33	40	30	28	39
kidSPAM4	32	39	31	41	44	43	38	46	38	40	38	47	24	36	27
kidSPAM3	33	37	29	42	41	34	45	44	45	37	37	42	27	34	26
kidSPAM2	34	38	28	43	42	33	43	43	43	38	41	39	29	35	25
ICTSPAM2	35	28	39	39	37	44	26	17	39	29	47	27	42	37	45
dalSPAM4	36	33	37	30	36	35	49	41	46	42	50	46	41	26	44
indSPAM3	37	41	36	40	45	31	37	49	33	45	35	43	40	42	37
pucSPAM0	38	27	32	36	33	38	34	21	37	12	25	35	31	39	36
indSPAM1	39	40	38	45	43	36	39	48	34	46	36	45	43	43	43
pucSPAM1	40	25	34	37	32	39	46	22	40	15	24	28	28	38	38
621SPAM2	41	24	42	47	23	45	51	25	51	30	26	25	26	24	42
pucSPAM2	42	46	33	34	31	37	47	28	41	34	30	49	35	49	35
ICTSPAM1	43	31	44	44	35	46	23	19	36	28	42	29	46	41	47
ICTSPAM3	44	43	45	48	47	48	50	47	50	44	45	48	47	46	48
ICTSPAM4	45	45	46	46	46	47	48	45	43	36	40	44	49	47	49
azeSPAM1	46	44	43	49	48	49	53	51	52	51	48	52	48	45	46
spamasas-v	-	-	-	24	24	27	10	5	1	35	28	13	-	-	-
popfile	-	-	-	21	16	18	20	33	9	22	12	14	-	-	-
tamSPAM4	-	-	-	-	-	-	15	20	17	13	34	33	-	-	-
tamSPAM3	-	-	-	19	17	20	18	24	22	19	39	31	-	-	-
indSPAM4	-	-	-	-	-	-	32	42	35	43	50	36	34	40	30
indSPAM2	-	-	-	-	-	-	44	52	32	53	49	53	44	43	41
azeSPAM2	-	-	-	-	-	-	52	53	53	50	50	51	45	48	34

Table 7: Summary Result Rankings

	trec05p-1/full			trec05p-1/s25			trec05p-1/s50			trec05p-1/h25			trec05p-1/h50		
Filters	hm%	sm%	lam%	hm%	sm%	lam%	hm%	sm%	lam%	hm%	sm%	lam%	hm%	sm%	lam%
621SPAM1	2.38	0.20	0.69	3.45	0.42	1.22	2.51	0.27	0.83	3.94	0.17	0.83	2.78	0.19	0.72
621SPAM2	55.06	1.07	10.32	54.53	1.34	11.33	54.80	0.86	9.32	58.82	1.39	12.42	57.10	1.18	11.20
621SPAM3	3.14	0.17	0.73	5.28	0.17	0.98	4.32	0.16	0.84	3.33	0.16	0.74	2.84	0.16	0.68
ICTSPAM1	5.69	20.85	11.19	3.01	16.37	7.23	6.05	13.75	9.20	15.15	3.74	7.69	10.64	8.51	9.52
ICTSPAM2	8.33	8.03	8.18	6.91	17.98	11.31	5.57	15.47	9.42	11.32	14.29	12.73	7.73	15.66	11.09
ICTSPAM3	14.10	28.22	20.26	14.42	23.67	18.61	13.50	27.17	19.44	13.68	26.99	19.49	12.51	27.21	18.78
ICTSPAM4	8.18	24.89	14.66	1.60	64.28	14.61	1.60	64.24	14.60	19.51	9.65	13.86	8.31	18.46	12.53
azeSPAM1	64.84	4.57	22.92	-	-	-	-	-	-	-	-	-	-	-	-
crmSPAM1	1.84	1.65	1.74	0.22	6.76	1.26	0.68	3.79	1.61	5.98	0.59	1.91	3.47	1.00	1.87
crmSPAM2	0.62	0.87	0.73	0.28	2.67	0.87	0.27	49.18	4.84	2.11	0.38	0.89	0.97	0.53	0.71
crmSPAM3	2.56	0.15	0.63	2.41	0.33	0.89	2.48	0.23	0.76	4.12	0.16	0.82	3.17	0.15	0.70
crmSPAM4	0.91	0.25	0.47	0.61	0.72	0.66	0.73	0.40	0.54	3.56	0.09	0.57	1.96	0.13	0.51
dalSPAM1	1.17	21.07	5.33	1.17	22.66	5.57	1.09	21.26	5.18	2.27	20.67	7.21	1.54	17.57	5.46
dalSPAM2	5.34	7.52	6.34	5.69	8.97	7.16	5.65	7.88	6.68	5.88	7.11	6.47	5.34	7.31	6.25
dalSPAM3	6.80	6.23	6.51	6.96	7.58	7.27	6.94	6.48	6.71	7.11	5.88	6.47	7.02	6.00	6.49
dalSPAM4	2.69	4.50	3.49	2.47	6.19	3.93	2.28	4.88	3.35	4.66	5.44	5.03	3.58	3.42	3.50
ijsSPAM1	0.25	0.93	0.48	-	-	-	-	-	-	0.32	1.02	0.57	-	-	-
ijsSPAM2	0.23	0.95	0.47	-	-	-	-	-	-	0.30	1.04	0.56	-	-	-
ijsSPAM3	0.26	0.97	0.51	-	-	-	-	-	-	0.38	1.11	0.65	-	-	-
ijsSPAM4	0.37	0.91	0.58	-	-	-	-	-	-	0.45	1.05	0.69	-	-	-
indSPAM1	0.82	15.16	3.70	0.70	21.48	4.21	0.75	17.58	3.86	1.75	11.02	4.49	1.20	13.11	4.10
indSPAM3	1.09	7.66	2.93	0.89	9.32	2.95	1.18	7.02	2.92	2.27	5.56	3.56	1.70	6.95	3.46
kidSPAM1	0.91	9.40	2.99	1.99	6.74	3.69	1.44	8.01	3.45	0.40	13.24	2.42	0.36	12.01	2.16
kidSPAM2	0.87	10.53	3.11	-	-	-	-	-	-	-	-	-	-	-	-
kidSPAM3	0.82	12.49	3.33	-	-	-	-	-	-	-	-	-	-	-	-
kidSPAM4	9.74	6.57	8.01	-	-	-	-	-	-	-	-	-	-	-	-
lbSPAM1	0.41	0.90	0.61	0.16	4.33	0.84	0.28	1.95	0.74	1.68	0.31	0.73	0.85	0.58	0.71
lbSPAM2	0.51	0.93	0.69	-	-	-	-	-	-	-	-	-	-	-	-
lbSPAM3	0.83	1.05	0.94	-	-	-	-	-	-	-	-	-	-	-	-
lbSPAM4	0.91	3.87	1.89	0.58	11.69	2.71	0.71	6.94	2.26	2.96	1.46	2.08	1.44	2.57	1.93
pucSPAM0	3.41	5.10	4.18	1.62	9.70	4.04	2.28	6.86	3.98	9.62	3.57	5.91	5.82	4.32	5.02
pucSPAM1	3.57	5.33	4.36	1.71	10.25	4.27	2.44	7.31	4.25	10.07	3.74	6.19	6.06	4.50	5.22
pucSPAM2	3.35	5.00	4.10	1.50	8.97	3.73	2.15	6.47	3.76	10.51	3.92	6.47	6.00	4.46	5.18
tamSPAM1	0.26	4.10	1.05	0.22	9.05	1.45	0.07	13.94	1.08	0.47	4.55	1.48	0.37	3.15	1.08
tamSPAM2	0.85	1.45	1.11	0.73	3.03	1.49	0.72	2.39	1.31	1.97	1.56	1.75	1.42	1.51	1.46
tamSPAM3	0.22	4.46	1.01	0.34	69.17	8.05	-	-	-	-	-	-	-	-	-
yorSPAM1	2.44	2.43	2.44	1.00	6.36	2.56	1.62	3.89	2.51	7.22	1.08	2.84	4.55	1.70	2.79
yorSPAM2	0.92	1.74	1.27	0.48	3.60	1.32	0.72	2.43	1.32	2.26	1.17	1.63	1.45	1.44	1.44
yorSPAM3	1.29	1.20	1.25	0.47	2.60	1.11	0.80	1.86	1.22	3.75	0.72	1.65	2.26	0.95	1.47
yorSPAM4	2.99	1.36	2.02	0.96	3.87	1.94	1.74	2.32	2.01	9.98	0.48	2.26	5.66	0.77	2.11

Table 8: Public Corpora Misclassification Summary

	trec05p-1/full			trec05p-1/s25			trec05p-1/s50			trec05p-1/h25			trec05p-1/h50		
Filters	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%	ROCA	h=.1	lam%
621SPAM1	0.044	3.63	0.69	0.091	4.14	1.22	0.048	2.72	0.83	0.070	6.65	0.83	0.054	5.34	0.72
621SPAM2	11.362	28.85	10.32	12.291	29.72	11.33	11.352	27.36	9.32	12.626	26.83	12.42	12.221	27.75	11.20
621SPAM3	0.060	7.02	0.73	0.085	6.72	0.98	0.061	7.07	0.84	0.068	7.58	0.74	0.058	6.28	0.68
ICTSPAM1	4.659	72.26	11.19	3.036	88.03	7.23	3.325	77.75	9.20	4.012	77.86	7.69	3.611	77.58	9.52
ICTSPAM2	2.643	79.51	8.18	4.571	89.15	11.31	2.741	85.34	9.42	6.140	95.18	12.73	3.777	83.79	11.09
ICTSPAM3	20.485	99.39	20.26	17.086	99.49	18.61	19.558	99.49	19.44	19.947	99.66	19.49	19.044	99.29	18.78
ICTSPAM4	10.952	98.44	14.66	27.891	97.00	14.61	27.506	96.29	14.60	10.821	99.58	13.86	8.995	98.67	12.53
azeSPAM1	28.887	99.50	22.92	-	-	-	-	-	-	-	-	-	-	-	-
crmSPAM1	0.169	10.53	1.74	0.236	9.64	1.26	0.194	10.58	1.61	0.383	43.87	1.91	0.219	18.37	1.87
crmSPAM2	0.122	4.52	0.73	0.343	5.23	0.87	41.915	50.14	4.84	0.097	22.25	0.89	0.067	7.59	0.71
crmSPAM3	0.042	2.63	0.63	0.051	2.96	0.89	0.044	2.64	0.76	0.066	6.42	0.82	0.051	2.11	0.70
crmSPAM4	0.049	1.96	0.47	0.089	1.90	0.66	0.055	1.36	0.54	0.069	11.63	0.57	0.059	3.26	0.51
dalSPAM1	2.348	99.75	5.33	2.662	99.73	5.57	2.183	99.76	5.18	2.997	99.47	7.21	2.026	99.50	5.46
dalSPAM2	1.674	41.92	6.34	1.970	56.39	7.16	1.827	49.81	6.68	1.713	41.31	6.47	1.694	40.60	6.25
dalSPAM3	1.491	41.00	6.51	1.814	51.35	7.27	1.635	47.24	6.71	1.453	40.80	6.47	1.459	38.51	6.49
dalSPAM4	1.370	76.58	3.49	1.854	85.10	3.93	1.430	82.06	3.35	2.087	82.63	5.03	1.217	71.00	3.50
ijsSPAM1	0.021	1.84	0.48	-	-	-	-	-	-	0.034	3.69	0.57	-	-	-
ijsSPAM2	0.019	1.78	0.47	-	-	-	-	-	-	0.031	3.15	0.56	-	-	-
ijsSPAM3	0.022	1.84	0.51	-	-	-	-	-	-	0.038	2.43	0.65	-	-	-
ijsSPAM4	0.025	2.22	0.58	-	-	-	-	-	-	0.041	4.03	0.69	-	-	-
indSPAM1	5.346	93.19	3.70	7.053	89.30	4.21	5.951	91.24	3.86	4.576	97.08	4.49	4.939	96.80	4.10
indSPAM3	2.822	97.35	2.93	2.844	97.56	2.95	2.471	98.18	2.92	3.210	98.53	3.56	3.012	97.59	3.46
kidSPAM1	1.463	34.93	2.99	1.589	55.96	3.69	1.546	46.36	3.45	1.812	26.90	2.42	1.586	27.99	2.16
kidSPAM2	4.544	91.65	3.11	-	-	-	-	-	-	-	-	-	-	-	-
kidSPAM3	4.167	90.62	3.33	-	-	-	-	-	-	-	-	-	-	-	-
kidSPAM4	3.990	93.74	8.01	-	-	-	-	-	-	-	-	-	-	-	-
lbSPAM1	0.039	4.56	0.61	0.092	5.71	0.84	0.054	4.75	0.74	0.081	14.26	0.73	0.056	10.10	0.71
lbSPAM2	0.037	5.19	0.69	-	-	-	-	-	-	-	-	-	-	-	-
lbSPAM3	0.122	22.38	0.94	-	-	-	-	-	-	-	-	-	-	-	-
lbSPAM4	0.238	22.94	1.89	0.588	20.53	2.71	0.347	21.61	2.26	0.332	46.84	2.08	0.261	33.31	1.93
pucSPAM0	2.083	59.71	4.18	2.200	65.46	4.04	2.083	61.97	3.98	2.600	59.58	5.91	2.314	56.50	5.02
pucSPAM1	2.185	52.58	4.36	2.623	48.07	4.27	2.367	51.10	4.25	2.618	49.60	6.19	2.409	55.12	5.22
pucSPAM2	1.967	51.28	4.10	1.788	54.12	3.73	1.853	52.57	3.76	3.274	52.25	6.47	2.358	54.32	5.18
tamSPAM1	0.164	6.92	1.05	0.483	12.33	1.45	1.004	11.97	1.08	0.234	10.71	1.48	0.123	6.10	1.08
tamSPAM2	0.178	27.38	1.11	0.268	15.61	1.49	0.225	16.81	1.31	0.326	60.91	1.75	0.323	54.26	1.46
tamSPAM3	0.183	7.64	1.01	22.663	71.24	8.05	-	-	-	-	-	-	-	-	-
yorSPAM1	2.032	87.24	2.44	3.234	92.91	2.56	2.369	89.66	2.51	3.292	91.80	2.84	2.564	89.86	2.79
yorSPAM2	0.457	34.21	1.27	0.420	24.46	1.32	0.426	29.56	1.32	0.669	38.53	1.63	0.530	35.53	1.44
yorSPAM3	0.861	62.13	1.25	1.176	56.54	1.11	1.025	64.74	1.22	1.382	72.84	1.65	1.082	70.10	1.47
yorSPAM4	0.688	84.92	2.02	0.586	78.32	1.94	0.537	84.19	2.01	1.975	90.82	2.26	1.117	89.26	2.11

Table 9: Public Corpora Summary Results

	Misclassified Spam (of 775 spams)							Misclassified Ham (of 6231 hams)							
	Automated	List	Newsletter	Phishing	Sex	Virus	**Total**	Automated	Commercial	Encrypted	Frequent	List	Newsletter	Personal	**Total**
621SPAM1	1	6	7	0	10	17	**41**	15	20	0	13	14	8	28	**98**
621SPAM2	1	9	7	3	15	18	**53**	20	15	18	29	15	9	48	**154**
621SPAM3	3	7	10	1	17	20	**58**	7	11	0	0	3	3	11	**35**
ICTSPAM1	11	21	14	5	83	6	**140**	6	6	0	6	27	9	9	**63**
ICTSPAM2	8	12	17	7	68	10	**122**	4	3	2	8	30	6	14	**67**
ICTSPAM3	5	17	11	1	114	2	**150**	14	29	45	56	154	64	50	**412**
ICTSPAM4	6	12	22	1	47	28	**116**	12	36	4	37	94	160	50	**393**
azeSPAM1	0	16	6	6	43	0	**71**	70	51	126	808	1938	255	360	**3608**
crmSPAM1	5	14	18	3	60	4	**104**	5	6	0	0	6	4	2	**23**
crmSPAM2	4	9	10	3	67	12	**105**	6	7	0	1	3	1	1	**19**
crmSPAM3	2	7	10	1	37	2	**59**	4	6	0	1	5	2	3	**21**
crmSPAM4	2	6	10	0	35	1	**54**	3	6	0	0	8	2	5	**24**
dalSPAM1	11	13	14	9	211	38	**296**	3	12	0	22	33	8	6	**84**
dalSPAM2	2	6	10	2	72	4	**96**	5	22	1	59	82	78	48	**295**
dalSPAM3	2	5	11	2	78	5	**103**	2	22	1	52	67	76	31	**251**
dalSPAM4	11	23	8	8	249	18	**317**	4	11	0	22	53	10	18	**118**
ijsSPAM1	3	9	4	1	66	5	**88**	6	6	0	1	6	2	1	**22**
ijsSPAM2	3	10	4	3	69	2	**91**	4	3	0	0	2	1	0	**10**
ijsSPAM3	2	7	3	0	69	5	**86**	9	10	0	1	12	3	2	**37**
ijsSPAM4	3	10	3	1	75	7	**99**	5	5	0	1	5	2	1	**19**
indSPAM1	5	18	19	6	251	19	**318**	4	5	0	10	34	55	6	**114**
indSPAM3	3	22	17	7	220	18	**287**	3	7	0	11	27	60	6	**114**
kidSPAM1	3	8	12	4	74	4	**105**	5	14	1	121	20	2	47	**210**
kidSPAM2	3	10	12	7	88	5	**125**	6	12	1	126	22	2	43	**212**
kidSPAM3	5	10	23	7	133	11	**189**	5	10	1	110	14	0	38	**178**
kidSPAM4	3	7	15	7	98	10	**140**	6	15	131	96	61	4	45	**358**
lbSPAM1	3	45	10	5	203	14	**280**	1	0	0	0	2	0	0	**3**
lbSPAM2	3	47	12	6	178	11	**257**	1	0	0	0	1	0	0	**2**
lbSPAM3	3	43	13	6	240	24	**329**	3	1	0	2	17	2	4	**29**
lbSPAM4	3	56	16	9	290	13	**387**	1	0	0	10	3	0	2	**16**
pucSPAM0	6	23	26	2	125	5	**187**	4	6	2	46	14	1	17	**90**
pucSPAM1	5	13	30	6	72	8	**134**	3	5	0	35	16	0	5	**64**
pucSPAM2	5	28	15	2	264	3	**317**	4	3	9	100	15	2	21	**154**
tamSPAM1	3	40	14	3	147	6	**213**	4	1	0	0	3	0	1	**9**
tamSPAM2	2	8	10	2	48	2	**72**	10	3	0	11	24	8	9	**65**
tamSPAM3	1	4	5	1	17	0	**28**	33	20	2	86	113	84	53	**392**
yorSPAM1	2	7	23	4	67	4	**107**	8	8	0	12	17	14	14	**74**
yorSPAM2	9	11	26	3	114	19	**182**	1	3	0	0	2	3	0	**9**
yorSPAM3	4	8	19	3	73	11	**118**	10	8	0	13	20	20	14	**85**
yorSPAM4	12	102	34	32	514	14	**708**	1	5	0	7	19	7	9	**48**

Table 10: Genre Classification of Misclassifications on S. B. Corpus

	Misclassified Spam (of 775 spams)							Misclassified Ham (of 6231 hams)							
	Automated	List	Newsletter	Phishing	Sex	Virus	**Total**	Automated	Commercial	Encrypted	Frequent	List	Newsletter	Personal	**Total**
ijsSPAM2	3	10	4	3	69	2	**91**	4	3	0	0	2	1	0	**10**
lbSPAM2	3	47	12	6	178	11	**257**	1	0	0	0	1	0	0	**2**
crmSPAM3	2	7	10	1	37	2	**59**	4	6	0	1	5	2	3	**21**
621SPAM1	1	6	7	0	10	17	**41**	15	20	0	13	14	8	28	**98**
tamSPAM1	3	40	14	3	147	6	**213**	4	1	0	0	3	0	1	**9**
yorSPAM2	9	11	26	3	114	19	**182**	1	3	0	0	2	3	0	**9**
dalSPAM4	11	23	8	8	249	18	**317**	4	11	0	22	53	10	18	**118**
kidSPAM1	3	8	12	4	74	4	**105**	5	14	1	121	20	2	47	**210**
pucSPAM2	5	28	15	2	264	3	**317**	4	3	9	100	15	2	21	**154**
ICTSPAM2	8	12	17	7	68	10	**122**	4	3	2	8	30	6	14	**67**
indSPAM3	3	22	17	7	220	18	**287**	3	7	0	11	27	60	6	**114**
azeSPAM1	0	16	6	6	43	0	**71**	70	51	126	808	1938	255	360	**3608**

Table 11: Genre Classification of Misclassifications on S. B. Corpus

The TREC 2005 Terabyte Track

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

Falk Scholer
RMIT
fscholer@cs.rmit.edu.au

Ian Soboroff
NIST
ian.soboroff@nist.gov

1 Introduction

The Terabyte Track explores how retrieval and evaluation techniques can scale to terabyte-sized collections, examining both efficiency and effectiveness issues. TREC 2005 is the second year for the track. The track was introduced as part of TREC 2004, with a single adhoc retrieval task. That year, 17 groups submitted 70 runs in total. This year, the track consisted of three experimental tasks: an adhoc retrieval task, an efficiency task and a named page finding task. 18 groups submitted runs to the adhoc retrieval task, 13 groups submitted runs to the efficiency task, and 13 groups submitted runs to the named page finding task. This report provides an overview of each task, summarizes the results and discusses directions for the future. Further background information on the development of the track can be found in last year's track report [4].

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. In 2005, the University of Glasgow took over the responsibility for distributing the collection. In 2004, the collection was distributed by CSIRO, Australia, who assisted in its creation.

3 The Adhoc Task

An adhoc task in TREC 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 the top 10,000 documents for that topic. For the 2005 task, NIST created and assessed 50 new topics. An example is provided in figure 1.

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

```
<top>
<num> Number: 756

<title> Volcanic Activity

<desc> Description:
Locations of volcanic activity which occurred within the present day
boundaries of the U.S. and its territories.

<narr> Narrative:
Relevant information would include when volcanic activity took place,
even millions of years ago, or, on the contrary, if it is a possible
future event.

</top>
```

Figure 1: *Adhoc Task Topic 756*

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 most experimental runs, participants could use any or all of the topic fields when creating queries from the topic statements. However, each 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. Groups could submit up to four runs.

The pools used to create the relevance judgments were based on the top 100 documents from two adhoc runs per group, along with two efficiency runs per group. This yielded an average of 906 documents judged per topic (min 347, max 1876). 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 the effectiveness measures, which require binary relevance judgments, the relevant and highly relevant documents were combined into a single "relevant" set.

In addition to the top 10,000 documents for each run, we collected details about the hardware and software configuration, 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.

Figure 2 provides an summary of the results obtained by the eight groups achieving the best results according to the bpref effectiveness measure [3]. When possible, we list two runs per group:

Group	Run	bpref	MAP	p@20	CPUs	Time (sec)
umass.allan	indri05AdmfS	0.4279	0.3886	0.5980	6	5162
	indri05Aql	0.3714	0.3252	0.5650	6	62
hummingbird.tomlinson	humT05xle	0.4264	0.3655	0.6230	1	50000
	humT05l	0.3659	0.3154	0.5800	1	5700
uglasgow.ounis	uogTB05SQE	0.4178	0.3755	0.6180	8	1000
uwaterloo.clarke	uwmtEwtaPt	0.3884	0.3451	0.5760	2	63
	uwmtEwtaD02t	0.2887	0.2173	0.4490	2	3
umelbourne.anh	MU05TBa2	0.3771	0.3218	0.5730	1	10
ntu.chen	NTUAH2	0.3760	0.3233	0.5630	1	734
	NTUAH1	0.3555	0.3023	0.5400	1	270
dublincityu.gurrin	DCU05AWTF	0.3603	0.3021	0.5600	5	120
tsinghua.ma	THUtb05SQWP1	0.3553	0.3032	0.5330	1	1800

Figure 2: *Adhoc Results (top eight groups by bpref)*

the run with the highest bpref and the run with the fastest time. The first two columns of the table identify the group and run. The next three columns provide the values of three standard effective measures for each run: bpref, mean average precision (MAP) and precision at 20 documents (p@20) [3]. The last two columns list the number of CPUs used to generate the run and the total query processing time. When the fastest and best runs are compared within groups, the trade-off between efficiency and effectiveness is apparent. This trade-off is further explored in the discussion of the efficiency results.

4 Efficiency Task

The efficiency task extends the adhoc task, providing a vehicle for discussing and comparing efficiency and scalability issues in IR systems by defining better methodology to determine query processing times. Nonetheless, the validity of direct comparisons between groups is limited by the range of hardware used, which varies from desktop PCs to supercomputers. Thus, participants are encouraged to compare techniques within their own systems or to compare the performance of their systems to that of public domain systems.

Ten days before the new topics were released for the adhoc task, NIST released a set of 50,000 efficiency test topics, which were extracted from the query logs of an operational search engine. Figure 3 provides some examples. The title fields from the new adhoc topics were seeded into this topic set, but were not distinguished in any way. Participating groups were required to process these topics automatically; manual runs were not permitted for this task.

Participants executed the entire topic set, reporting the top-20 results for each query and the total query processing time for the full set. Query processing time included the time to read the topics and write the final submission file. The processing of topics was required to proceed sequentially, in the order the topics appeared in the topic file. To measure effectiveness, the results corresponding to the adhoc topics were extracted and added into the evaluation pool for the adhoc task. Since the efficiency runs returned only the top 20 documents per topic, they did not substantially increase the pool size.

```
7550:yahoo
7551:mendocino and venues
7552:creative launcher
7553:volcanic activity
7554:shorecrest
7555:lazy boy
7556:los bukis deseo download free
7557:online surveys
7558:wholesale concert tickets
```

Figure 3: *Efficiency Task Topics 7550 to 7558*

Group	Run	p@20	CPUs	Time (sec)	US$ Cost
uwaterloo.clarke	uwmtEwtePTP	0.5780	2	54701	3800
	uwmtEwteD10	0.3900	2	1371	3800
umelbourne.anh	MU05TBy1	0.5620	8	2145	6000
	MU05TBy3	0.5550	8	1201	6000
hummingbird.tomlinson	humTE05i4ld	0.5490	1	219354	5000
	humTE05i5	0.4460	1	39506	5000
umass.allan	indri05Eql	0.5490	1	71700	1500
	indri05EqlD	0.5490	6	24720	9000
rmit.scholer	zetdir	0.5410	1	11565	1200
	zetdist	0.5300	8	2901	6000
dublincityu.gurrin	DCU05DISTWTF	0.5290	5	48375	13125
	DCU05WTFQ	0.4660	1	17730	2625
ntu.chen	NTUET2	0.5180	1	186900	2400
	NTUET1	0.5150	1	183200	2400
upisa.attardi	pisaEff2	0.4350	23	12898	10000
	pisaEff4	0.3420	23	7158	10000

Figure 4: *Efficiency Results (top eight groups by p@20)*

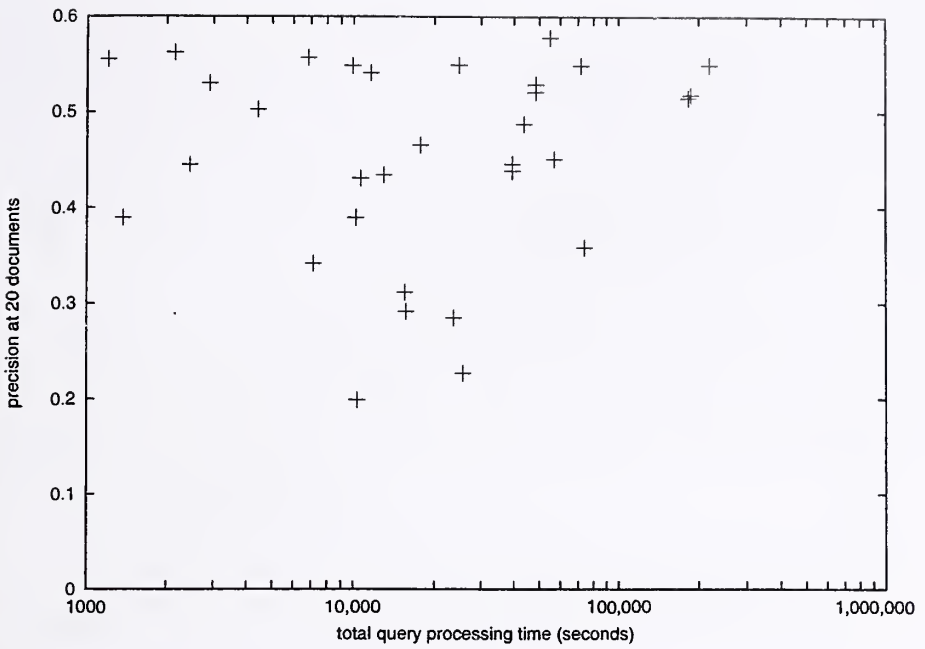

Figure 5: *Efficiency vs. Effectiveness*

Figure 4 summarizes the results for the eight groups achieving the best results, based on p@20. Once again, the figure lists both the best and fastest run from each group. In addition to the query processing times and number of CPUs, the table also includes the estimate of total hardware cost provided by each participating group.

To illustrate the range of results seen within the track, and to provide a sense of the trade-offs between efficiency and effectiveness, figure 5 plots p@20 against total query processing time for all 32 runs submitted to the efficiency track. Note that a log scale is used to plot query processing times. The range in both dimensions is quite dramatic.

The results plotted in figure 5 were generated on a variety of hardware platforms, with different costs and configurations. To adjust for these differences, we attempted two crude normalizations. Figure 6 plots p@20 against total query processing time, normalized by the number of CPUs. The normalization was achieved simply by multiplying the time by the number of CPUs. Figure 7 plots p@20 against total query processing time, normalized by hardware cost, with the times adjusted to a typical uniprocessor server machine costing $2,000. In this case, the normalization consisted of multiplying the time by the cost and dividing by 2,000. Both normalizations have the effect of moving the points sharply away from the upper left-hand corner, making the trade-offs in this area more apparent.

5 Named Page Finding Task

Named page finding is a navigational search task, where a user is looking for a particular resource. In comparison to an adhoc task, a topic for a named page finding task usually has one answer: the resource specified in the query. The objective of the task, therefore, is to find a particular page in

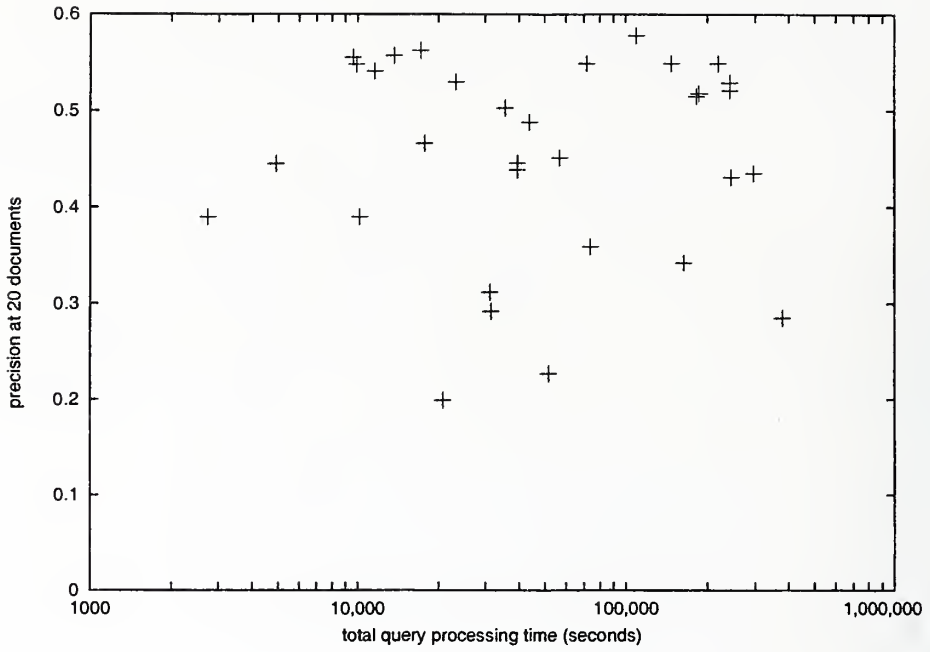

Figure 6: *Efficiency vs. Effectiveness (normalized by number of CPUs)*

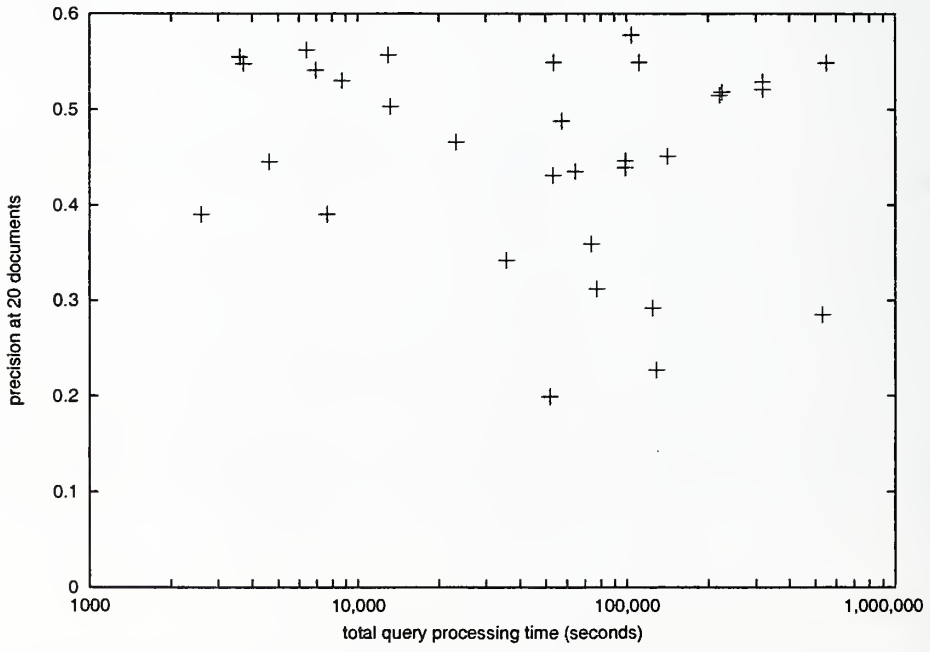

Figure 7: *Efficiency vs. Effectiveness (normalized by hardware cost)*

the GOV2 collection, given a topic that describes it. For example, the query "`fmcsa technical support`" would be satisfied by the Federal Motor Carrier Administration's technical support page.

Named page topics were created by envisaging a bookmark recovery process. Participants were presented with a randomly selected page from the GOV2 collection. If the page had any identifiable features, and looked like something that a real user might want to read, remember, and re-find at a later point in time, the participant was asked to write a query. Specifically, the task was: "write a query to retrieve the page, as if you had seen it 15 minutes ago, and then lost it". This process resulted in an initial list of 272 queries, each with one corresponding target page. After manual inspection, 20 of these were discarded because they were "topic finding" in nature. We believe that collection browsing facilities would have made topic creation far easier than merely viewing successive random pages, but this facility was not available at the time of topic creation.

Although named page topics are typically assumed to have a single correct answer, the topic creation process outlined above raises two issues: first, there may be exact duplicates of the target document in the collection; and second, the target may not be specified clearly enough to rule out topically similar pages as plausible answers.

The first issue was resolved by searching for near-exact duplicates of the target pages within the GOV2 collection. This was done with the DECO system [1], which uses a lossless fingerprinting technique for the detection of duplicate documents. Pools formed from the top 1000 answers from all 42 runs that were submitted for this year's task (around 1.5 million unique documents) were searched. All near-exact duplicates were included in the qrels file.

In the context of past TREC Web Tracks, the second issue was sometimes resolved by requiring the creation of "omniscient" queries; that is, each query is tested and, if it retrieves similar (but not identical) documents in the collection that could also be considered to be plausible answers, it is discarded. However, discarding such queries distances the experimental process from a real-world web search task: a user generally does not know in advance if a named page query is specific enough to only identify a single resource. For the Terabyte Track, we therefore chose to retain such queries, and treated them as having a single "correct" answer. As a result of the change in methodology, we expected that the named page finding task would be harder than previously experienced.

Of the 252 topics used for the named page finding task, 187 have a single relevant answer (that is, there are no exact duplicates that match the canonical named page in the answer pool). However, some pages repeat often in the collection (the highest number of duplicate answer pages were identified for topic 778, with 4525 repeats).

Figure 8 summaries 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 top eight groups 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). While reasonable results can be achieved without exploiting these web-related document characteristics, most of the top runs incorporate one or more of them.

Group	Run	MRR	% Top 10	% Not Found	CPUs	Time (sec)	Links?	Anchors?	Structure?
tsinghua.ma	THUtb05npW15	0.463	61.5	17.9	1	5400	N	Y	N
umass.allan	indri05Nmpsd	0.441	58.3	17.1	6	16200	Y	Y	Y
uglasgow.ounis	uogNP05baseN	0.401	54.8	19.8	64	3840	N	Y	Y
ntu.chen	NTUNF3	0.388	51.2	19.4	1	46200	N	Y	Y
hummingbird.tomlinson	humTN05pl	0.378	50.0	19.8	1	14000	N	N	Y
uwaterloo.clarke	uwmtEwtnpP	0.366	50.8	20.6	2	944	N	N	N
uamsterdam.mueller	UAmsT05n3SUM	0.365	48.8	22.6	2	13239	N	Y	N
yorku.huang	york05tNa1	0.329	44.4	25.4	1	10365	N	N	N

Figure 8: *Named Page Finding Results (top 8 groups by MRR)*

6 The Limits of Pooling

Aside from scaling TREC and research retrieval systems, a primary goal of the terabyte track is to determine if the Cranfield paradigm of evaluation scales to very large collections, and to propose alternatives if it doesn't. In the discussions which led to the current terabyte track, the hypothesis was that in very large collections we would not be able to find a good sample of relevant documents using the pooling method. If judgments are not sufficiently complete, then runs which were not pooled will retrieve unjudged documents, making those runs difficult to measure. Depending on the run and the reason that the judgments are incomplete, this can result in a biased test collection. According to MAP, unjudged documents are assumed irrelevant and a run may score lower than it should. The bpref measure can give artificially high scores to a run if they do not retrieve sufficient judged irrelevant documents.

One method for determining if pooling results in insufficiently complete judgments is to remove a group's pooled runs from the pools, and measure their runs as if they had not contributed their unique relevant documents. This test does reveal that the terabyte collections should be used with some caution. For 2004, the mean difference in scores across runs when the run's group is held out is 9.6% in MAP, and the maximum is 45.5%; on the other hand, this was the first year of the track, and overall system effectiveness was not very good. This year, the mean difference is 3.9% and the maximum is 17.7%, much more reasonable but still somewhat higher than we see in newswire.

Another approach is to examine the documents to see if they seem to be biased towards any particular retrieval approach. We have found that the relevant documents in both terabyte collections have a very high occurrence of the title words from the topics, and that this occurrence is much higher than we see in news collections for ad hoc retrieval. More formally, the *titlestat* measure for a set of documents D is defined to be the percentage of documents in D that contain a title word, computed as follows. For each word in the title of a topic that is not a stop word, calculate the percentage of the set D that contain the word, normalized by the maximum possible

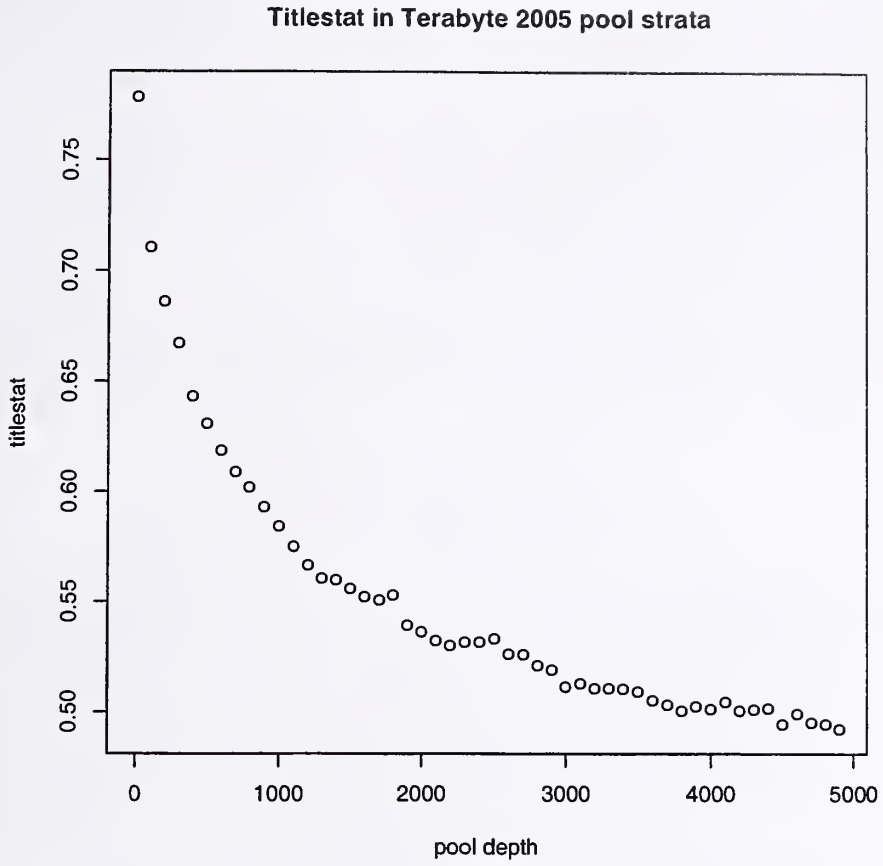

Figure 9: Titlestat in 100-document strata of the 2005 terabyte pools.

percentage.[1] For a topic, this is averaged over all words in the title, and for a collection, averaged over all topics. A maximum of 1.0 occurs when all documents in D contain all topic title words; 0.0 means that no documents contain a title word at all. Titlestat can be thought of as the occurrence of the average title word in the document set.

Titlestat can be measured for any set of documents. For the relevant documents (*titlestat_rel*)in the terabyte collections, we obtain 0.889 for the 2004 collection and 0.898 for 2005. In contrast, the TREC-8 ad hoc collection (TREC CDs 4 and 5 less the Congressional Register) has a *titlestat_rel* of 0.688. For the WT10g web collection, the TREC-9 ad hoc task relevant documents have a *titlestat_rel* of 0.795, and TREC-10 is 0.761.

Why are the terabyte titlestats so high? We feel that this is directly due to the size of the collection. In nearly any TREC collection, the top ranked documents are going to reflect the title words, since (a) title-only runs are often required, (b) even if they are not required, the title words are often used as part of any query, and (c) most query expansion will still weight the original query

[1] In rare cases, a title word will have a collection frequency smaller than $|D|$.

terms (i.e., the title words) highly. So it's certainly expected that the top ranks will be dominated by documents that contain the title words. For GOV2, the collection is so large that the title words have enormous collection frequency compared to the depth of the assessment pool. The result of this is that the pools are completely filled with title-word documents, and documents without title words are simply not judged.

Figure 9 illustrates this phenomenon using the titlestat of the pools, rather than of the judged relevant documents. The first point at $x = 0$ is the titlestat of the pool from depth 1-100, the pool depth used this year (0.778). In contrast, the titlestat of the TREC-8 pools is 0.429. Subsequent points in the graph show the titlestat of the pool from depth 101-200, 201-300, and so forth. Each pool is cumulative with respect to duplicates, meaning that if a document was pooled at a shallower depth it is not included in a deeper pool stratum. In order to get to lower titlestat depths, we would have had to pool very deep indeed. In any event, these titlestats indicate that the pools are heavily biased towards documents containing the title words, and may not fairly measure runs which do not use the title words in their query.

7 The Future of the Terabyte Track

Our analysis of title-word occurrence within the terabyte pools and relevance judgments indicates that the terabyte collections may be biased towards title-only runs. This is a serious concern for a TREC collection, and for the 2006 adhoc task we intend to pursue several strategies to build a more reusable test collection. In part, greater emphasis will be placed on the submission of manual runs, expanding the variety of relevant documents in the pools to include more documents that contain few or none of the query terms and increasing the re-usability of the collection. Users of the 2004 and 2005 collections should be very cautious. We recommend the use of multiple effectiveness measures (such as MAP and bpref) and careful attention to the number of retrieved unjudged documents.

In addition, the evaluation procedure may be modified to reduce the influence of *content-equivalent* documents in the collection. Using the 2004 topics as a case study, Bernstein and Zobel [2] present methods for identifying these near-duplicate documents and discover a surprisingly high level of inconsistency in their judging. Moreover, these near duplicates represent up to 45% of the relevant documents for given topics. This inconsistency and redundancy has a substantial impact on effectiveness measures, which we intend to address in the definition of the 2006 task.

Along with the adhoc task, we plan to run a second year of the efficiency and named page finding tasks, allowing groups to refine and test methods developed this year. In the case of the efficiency task, we are developing a detailed query execution procedure, with the hope of allowing more meaningful comparisons between systems.

Planning for 2006 is an ongoing process. As our planning progresses, it is possible that we may add an efficiency aspect to the named page finding task, and a "snippet retrieval" aspect to the adhoc retrieval task. A substantial expansion of the test collection remains a long-term goal, if the track continues in the future.

Acknowledgments

We thank Nick Craswell, David Hawking and others at CSIRO for their assistance with the creation of the GOV2 collection, and Doug Oard who hosted our Web crawler at the University of Maryland. The members of the Search Engine Group at RMIT University helped in the creation of the named page topics for this year's Terabyte Track. Finally, we thank Yaniv Bernstein, who kindly made his DECO software available for the identification of duplicate answer pages.

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] Yaniv Bernstein and Justin Zobel. Redundant documents and search effectiveness. In *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management*, pages 736–743, Bremen, Germany, 2005.

[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`.

Text REtrieval Conference (TREC)
...to encourage research information retrieval
from large text collections
TREC 2005
Conference
Proceeding
SP 500-266
NIST

NIST Technical Publications

Periodical

Journal of Research of the National Institute of Standards and Technology—Reports NIST research and development in metrology and related fields of physical science, engineering, applied mathematics, statistics, biotechnology, and information technology. Papers cover a broad range of subjects, with major emphasis on measurement methodology and 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