Overview of TREC 2006

Sponsored by:
NIST, DTO

Ellen Voorhees

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


TREC
15th
Anniversary
Text REtrieval Conference (TREC)

TREC 2006 Program Committee

Ellen Voorhees, chair
James Allan
Chris Buckley
Gord Cormack
Sue Dumais
Donna Harman
Bill Hersh
David Lewis
John Prager
Steve Robertson
Mark Sanderson
Ian Soboroff
Karen Sparck Jones
Richard Tong
Ross Wilkinson

TREC 2006 Track Coordinators

Blog: Ounis, de Rijke, Macdonald, Mishne, Soboroff

Enterprise: Nick Craswell, Arjen de Vries, Ian Soboroff

Genomics: William Hersh

Legal: Jason Baron, David Lewis, Doug Oard

Question Answering: Hoa Dang, Jimmy Lin, Diane Kelly

Spam: Gordon Cormack

Terabyte: Stefan Büttcher,Charles Clarke,Ian Soboroff

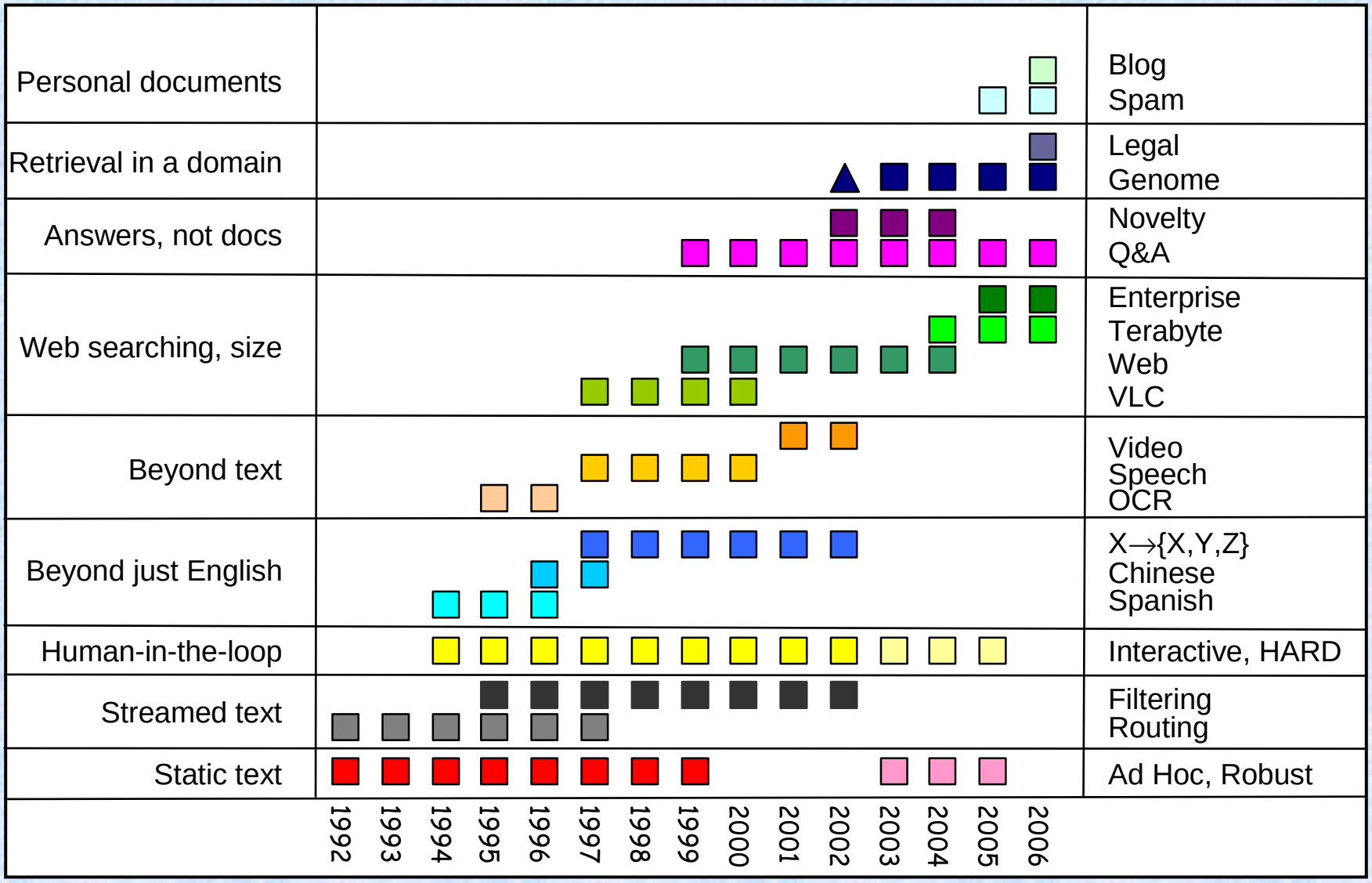
TREC 2006 Participants

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

TREC Goals

- To increase research in information retrieval based on large-scale collections
- To provide an open forum for exchange of research ideas to increase communication among academia, industry, and government
- To facilitate technology transfer between research labs and commercial products
- To improve evaluation methodologies and measures for information retrieval
- To create a series of test collections covering different aspects of information retrieval

The TREC Tracks



Common Terminology

- "Document" broadly interpreted,
  - for example
    - email message in enterprise, spam tracks
    - blog posting plus comments in blog track
- Classical IR tasks
    - ad hoc search: collection known; new queries
    - routing/filtering: standing queries; streaming document set
    - known-item search: find partially remembered specific document
    - question answering

TREC 2006 Themes

- Explore broader information contexts
  - different document genres
    - newswire (QA); web (terabyte)
    - blogs; email (enterprise, spam); corporate repositories (legal, enterprise); scientific reports (genomics, legal)
  - different tasks
    - ad hoc (terabyte, enterprise discussion, legal, genomics); known-item (terabyte); classification (spam)
    - specific responses (QA, genomics, enterprise expert); opinion finding (blog)

TREC 2006 Themes

- Construct evaluation methodologies
  - fair comparisons on massive data sets (terabyte, legal)
  - quality of a specific response (QA, genomics)
  - evaluation strategies balancing realism and privacy protection (spam, enterprise)
  - distributed efficiency benchmarking (terabyte)

Terabyte Track

- Motivations
  - investigate evaluation methodology for collections substantially larger than existing TREC collections
  - provide test collection for exploring system issues related to size
- Tasks
  - traditional ad hoc retrieval task
  - named page finding task
  - efficiency task
    - systems required to report various timing and resource statistics

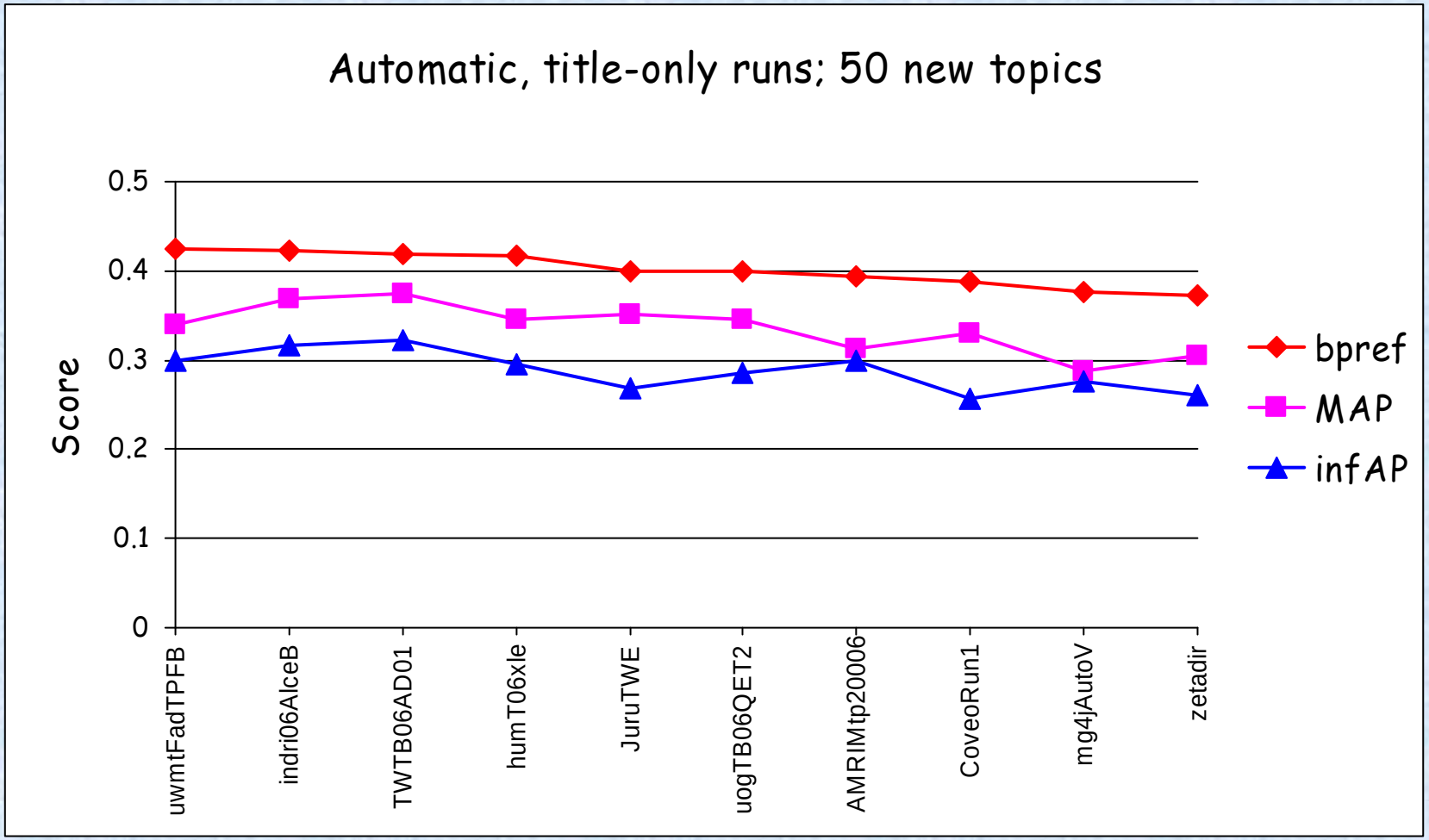
Terabyte Collection

- Documents
  - ~ 25,000,000 web documents (426 gb)
  - spidered in early 2004 from .gov domain
  - includes text from pdf, word, etc. files
- Topics
  - 50 new information-seeking topics for ad hoc
  - 181 named page queries contributed by participants
  - 100,000 queries from web logs for efficiency
- Judgments
  - new pooling methodology in ad hoc to address bias
  - duplicate detection using DECO

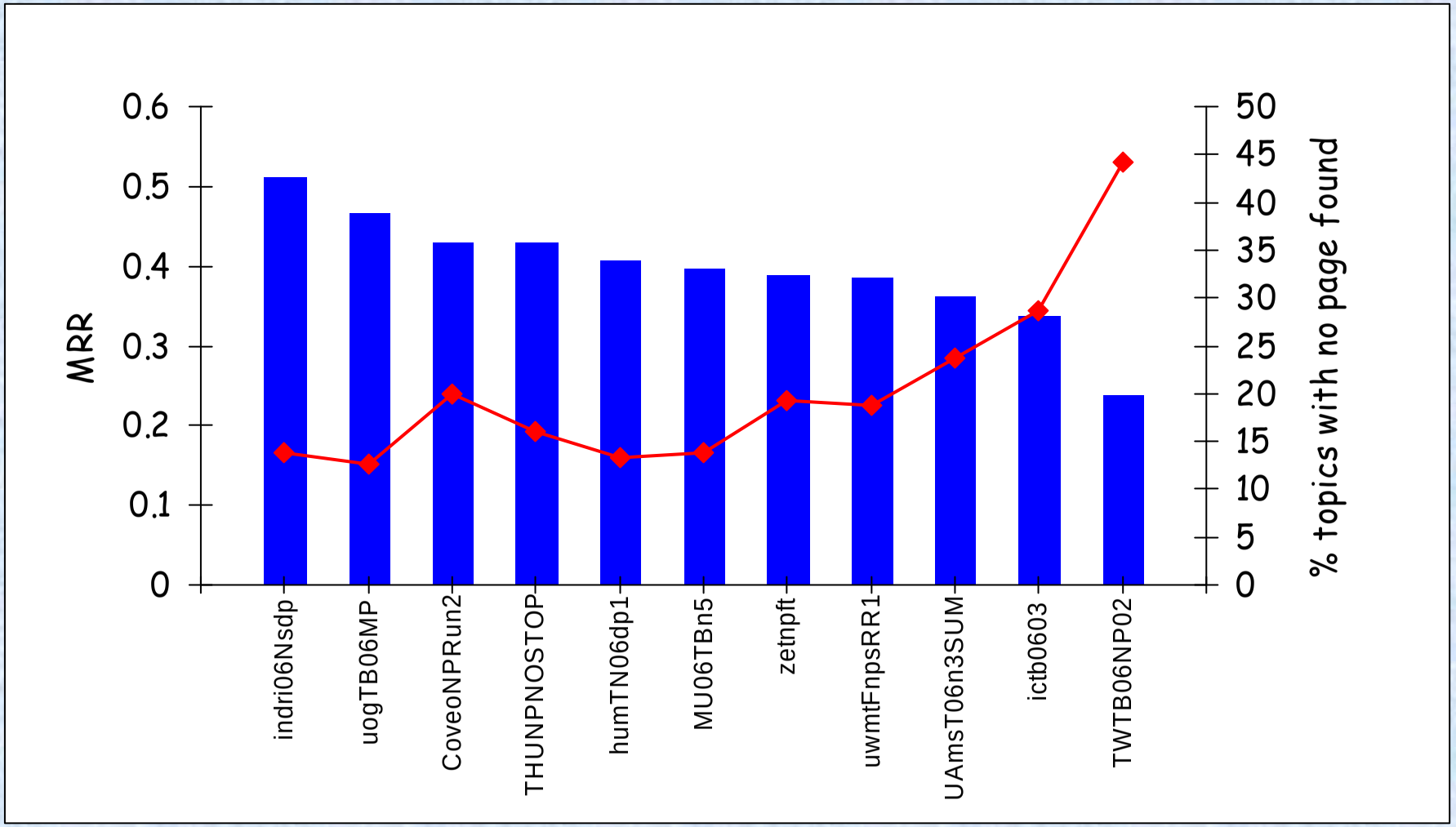
Ad Hoc Task

- Emphasis on exploring pooling strategies
  - task guidelines strongly encouraged manual runs to help pools
    - [unspecified] prize offered for most unique rels
  - judgment budget allocated among 3 pools
    - traditional pools to depth 50; used for scoring runs by trec_eval
    - traditional pools starting at rank 400; not used for scoring
    - sample of documents per topic such that number judged topic-dependent; used for scoring runs by inferred AP [Yilmaz & Aslam]

Ad Hoc Task Results



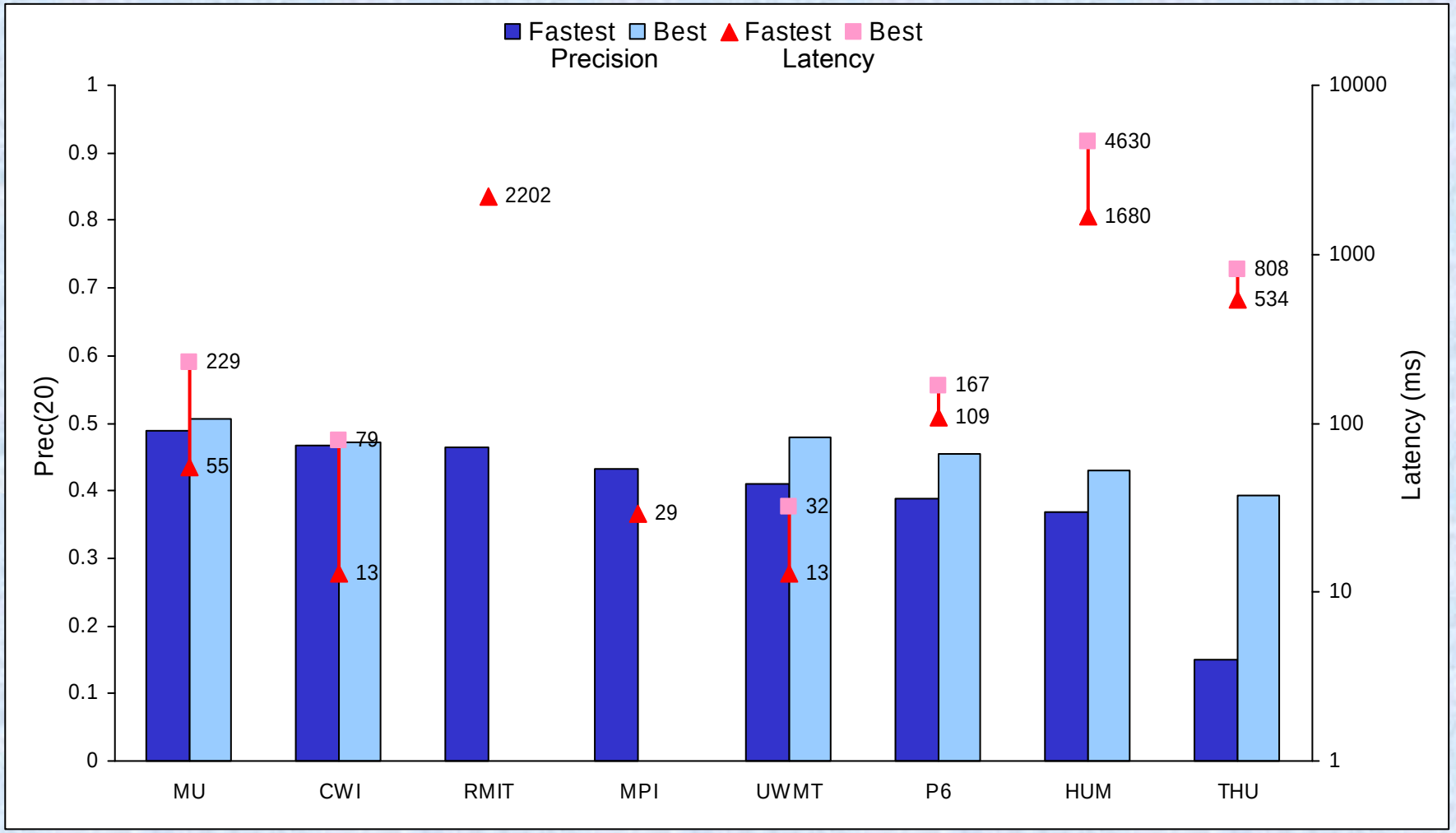
Named Page Finding Results



Efficiency Task

- Goal: compare efficiency/scalability despite different hardware
  - 100,000 queries vetted for hits in corpus and seeded with queries from other tasks
  - queries divided into 4 streams; within-stream queries required to be processed serially; between streams could use arbitrary interleaving
  - participants also report efficiency measures for running open source retrieval system with known characteristics for normalization
  - P(20) or MRR measured for seeded queries

Efficiency Task Results



Question Answering Track

- Goal: return answers, not document lists
- Tasks:
  - define a target by answering a series of factoid and list questions about that target, plus returning other info not covered by previous questions
  - complex interactive question answering (ciQA)
- AQUAINT document collection source of answers for all tasks
  - 3 GB text; approx. 1,033,000 newswire articles

Question Series

185	Iditarod race		
	185.1	FACT	In what city does the Iditarod start?
	185.2	FACT	In what city does the Iditarod end?
	185.3	FACT	In what month is it held?
	184.4	FACT	Who is the founder of the Iditarod?
	185.5	LIST	Name people who have run the Iditarod
	185.6	FACT	How many miles long is the Iditarod?
	185.7	FACT	What is the record time in which the Iditarod was won?
	185.8	LIST	Which companies have sponsored the Iditarod?
	185.9	Other	

75 series in test set with 6-9 questions per series

- 19 People
- 19 Organizations
- 18 Things
- 19 Events

- 403 total factoid questions
- 89 total list questions
- 75 total "other" questions

Main task

- Similar to previous 2 years, but:
  - questions required to be answered within particular time frame
    - present tense latest in corpus
    - past tense implicit default reference to timeframe of series
  - factoid questions relatively less important in final score
  - conscious attempt to make questions somewhat more difficult
    - temporal processing

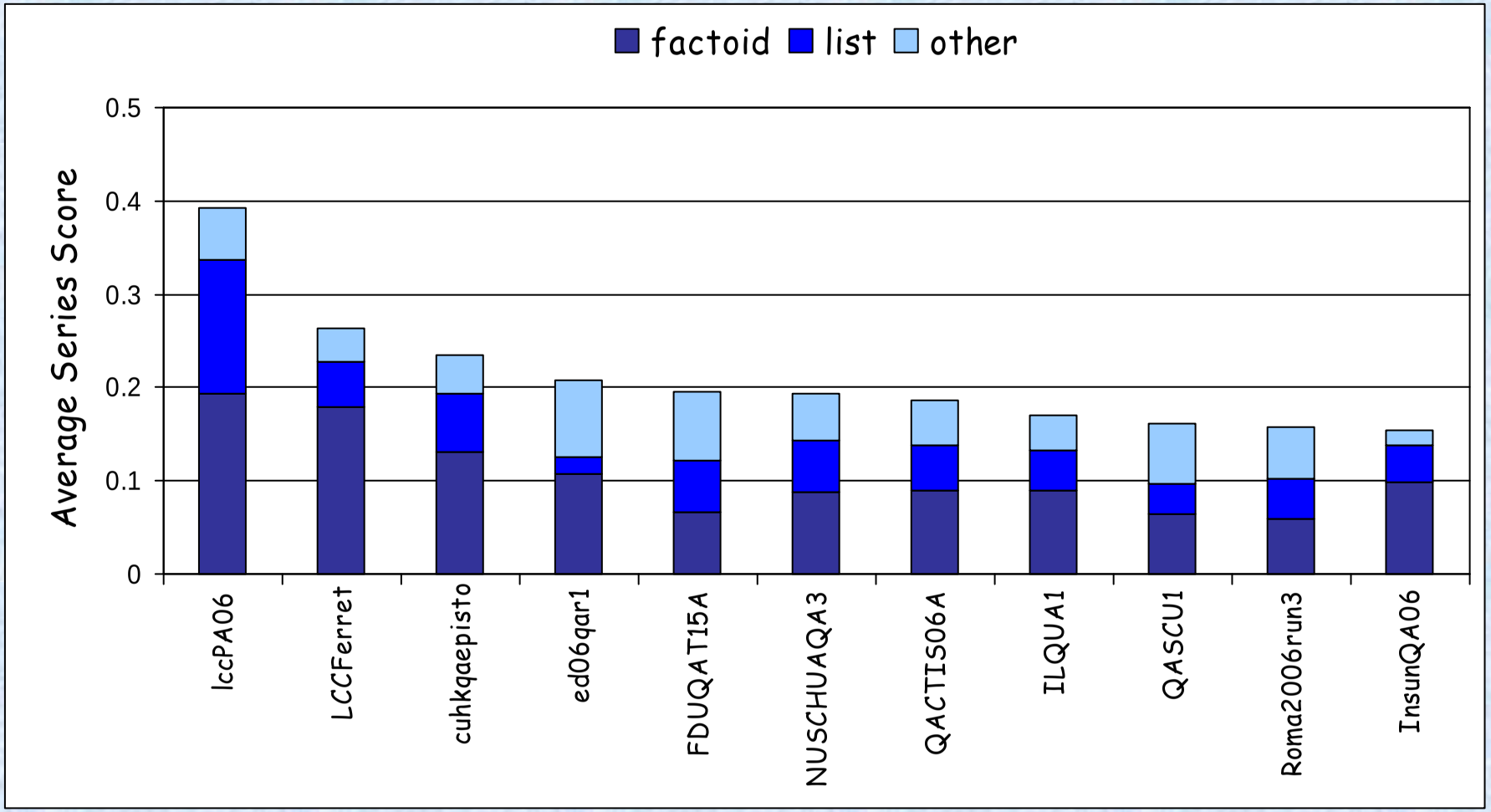
Series Score

- Score a series using weighted average of components

  Score = 1/3FactoidScore+1/3ListScore+1/3OtherScore

- Component score is mean of scores for questions of that type in given series
  - **FactoidScore**: average accuracy. Individual question has score of 1 or 0
  - **ListScore**: average F measure score. Recall & precision of response based on entire set of known answers.
  - **OtherScore**: F(β=3) score for that series' Other question, calculated using "nugget" evaluation

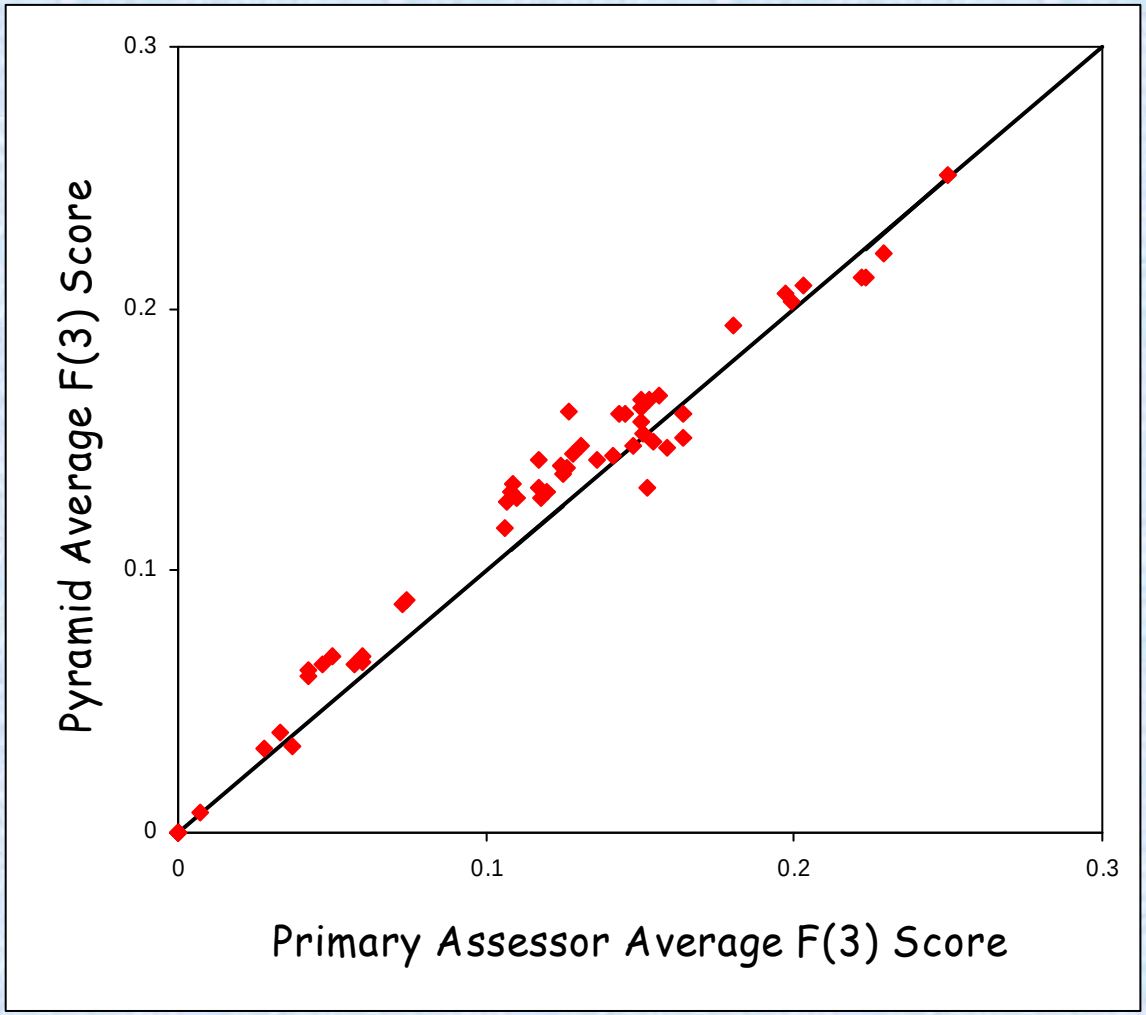
Series Task Results



Pyramid Nugget Scoring

- Variant of nugget scoring for 'Others'
  - suggested by Lin and Demner-Fushman [2006]
  - vital/okay distinction by single assessor is major contributor to instability of the evaluation; thus have multiple judges "vote" on nuggets, and use percentage voting for a nugget as nugget's weight
  - each QA assessor given list of nuggets created by primary assessor for series and asked simply to say vital/okay/NaN
  - pyramid of votes does not represent any single user, but does increase stability and average series scores highly correlated

Single vs. Pyramid Nugget Scores



Complex Interactive QA

- Goals:
  - investigate richer user contexts within QA
  - have (limited) actual interaction with user
- Task inspired by TREC 2006 relationship QA task and HARD track
  - "essay" questions
  - interaction forms allowed participants to solicit information from assessor (surrogate user)

Complex Questions

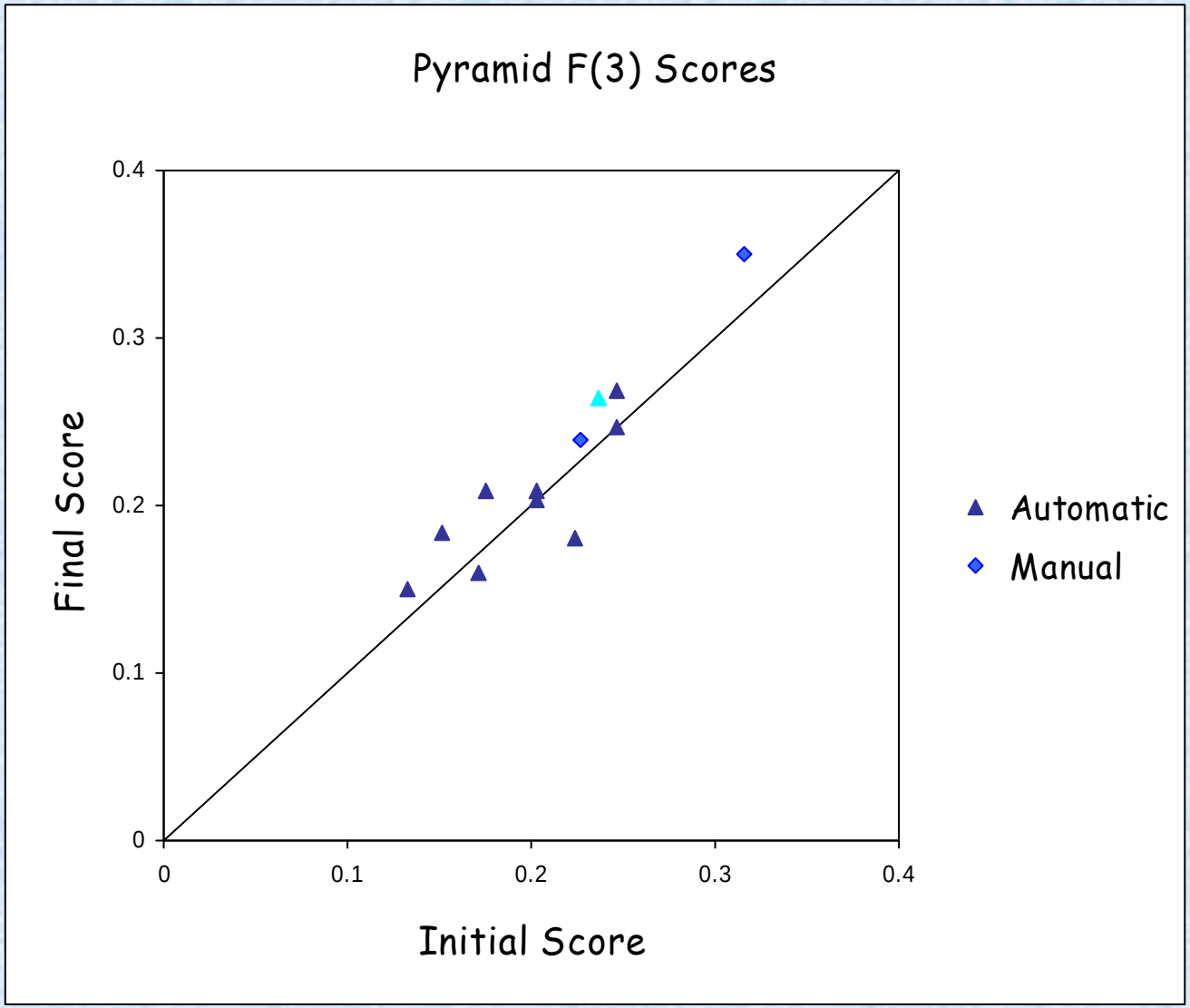
- Questions taken from relationship type identified in AQUAINT pilot
  - question formed from a relationship template
  - also included narrative giving more details

What evidence is there for transport of [goods] from [entity] to [entity]?
What [financial relationship] exists between [entity] and [entity]?
What [organizational ties] exist between [entity] and [entity]?
What [familial ties] exist between [entity] and [entity]?
What [common interests] exist between [entity] and [entity]?
What influence/effect does [entity] have on/in [entity]?
What is the position of [entity] with respect to [issue]?
Is there evidence to support the involvement of [entity] in [event/entity]?

ciQA Protocol

- Perform baseline runs
- Receive interaction form responses
  - interaction forms web forms that ask assessor for more information
  - content is participant's choice, but assessor spends at most 3 minutes/topic/form
- Perform additional (non-baseline) runs exploiting additional info

Relationship Task Results



Genomics Track

- Track motivation: explore information use within a specific domain
  - focus on person experienced in the domain
- 2006 task
  - single task with elements of both IR and IE
  - instance-finding, QA task where unit of retrieval is a (sub) paragraph

Genomics Track Task

- Documents
  - full-text journal articles provided through Highwire Press
  - associated metadata (eg MEDLINE record) available
  - 162,259 articles from 49 journals; about 12.3GB HTML
- Topics
  - 28 well-formed questions derived from topics used in 2005 (two questions discarded as had no relevant passages)
  - topics based on 4 generic topic type templates and instantiated from real user requests
    - e.g., *What is the role of DRD4 in alcoholism?*
      How do HMG and HMGB1 interact in hepatitis?
- System response
  - ranked list of up to 1000 passages (pieces of paragraphs)
  - each passage a contribution to the answer

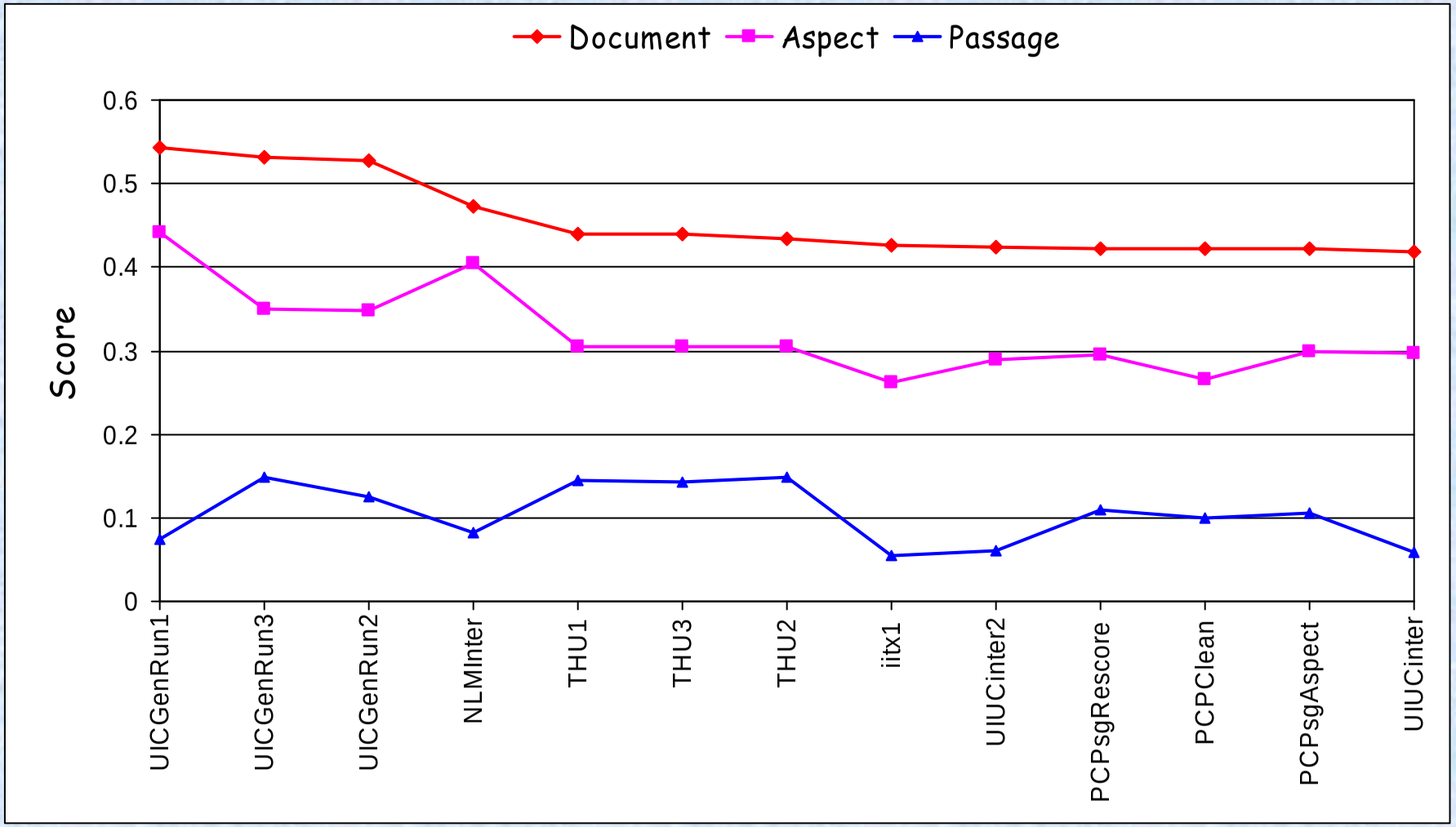
Task Evaluation

- Relevance judging
  - pools built from standardized paragraphs by mapping retrieved passage to its unique standard
  - judged by domain experts using 3-way judgments: not/possibly/definitely relevant
  - assessor marked contiguous span in paragraph as answer
  - answers assigned a set of MeSH terms

Task Evaluation

- Scoring
  - document:
    - standard ad hoc retrieval task (MAP)
    - doc is relevant iff it contains a relevant passage
    - collapse system ranking so doc appears just once
  - aspect:
    - use MeSH term sets to define aspects
    - retrieved passage that overlaps with marked answer assigned aspect(s) of that answer
  - passage
    - calculate fraction of answers (in characters) that are contained in retrieved passages

Genomics Track Results



Enterprise Track

- Goal: investigate enterprise search, searching the data of an organization to complete some task
  - find-an-expert task
  - ad hoc search for email messages that discuss reasons pro/con for a particular past decision
- Document set
  - crawl of W3C public web as of June 2004
  - ~300,000 documents divided into five major areas: discussion lists (email), code, web, people, other

Search-for-Experts Task

- Motivation
  - exploit corporate data to determine who are experts on a given topic
- Task
  - given a topic area, return a ranked-list of people; also return supporting documents
  - people are represented as an id from a canonical list of people in the corpus
  - evaluate as standard ad hoc retrieval task where a person is relevant if he/she is indeed an expert [and expertise is documented]
  - topics and judgments constructed by participants

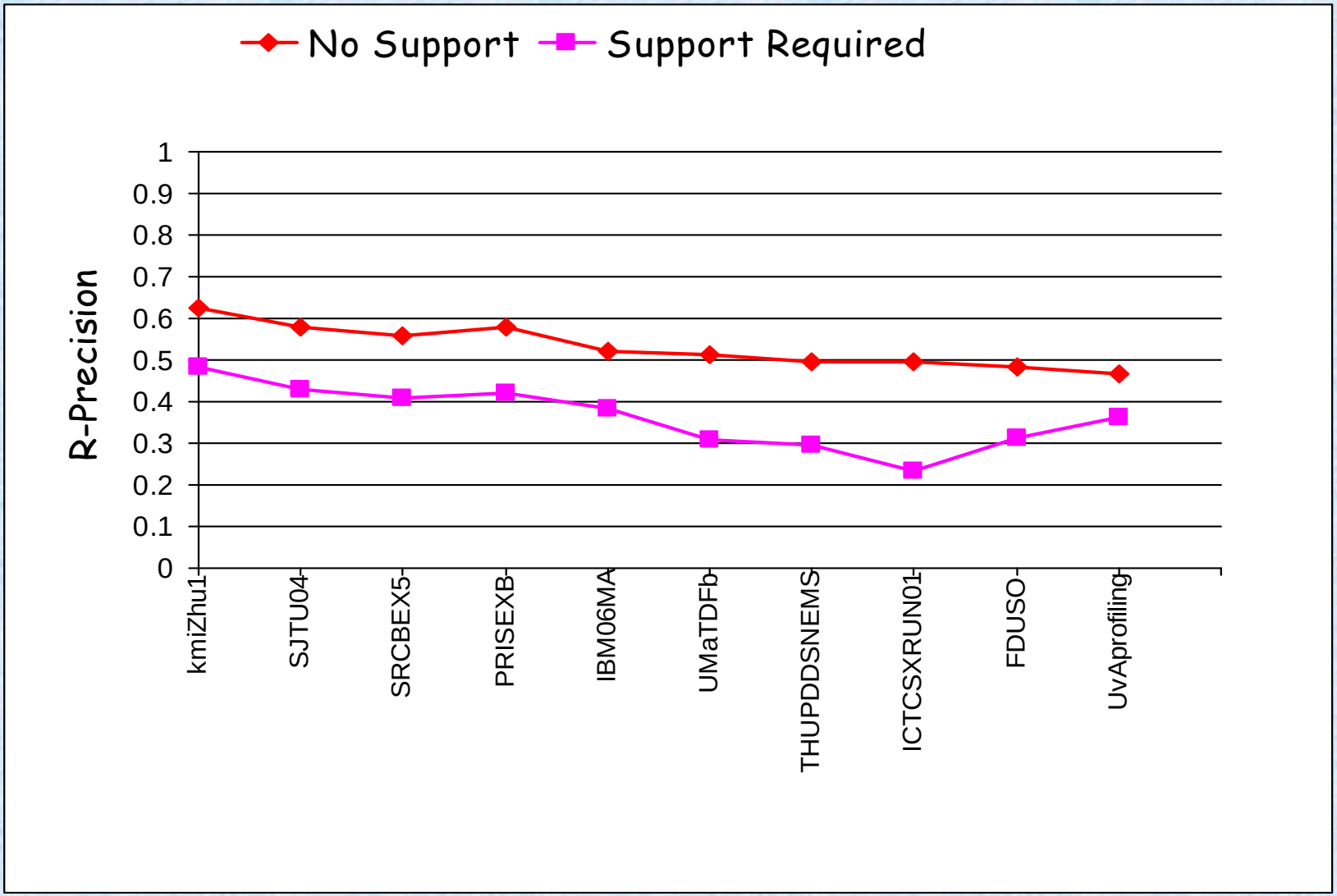
Expert Search Topic

<title>
W3C translation policy

<description>
I want to find people in charge or knowledgeable about W3C translation policies and procedures

<narrative>
I am interested in translating some W3C document into my native language. I want to find experts on W3C translation policies and procedures. Experts are those persons at W3C who are in charge of the translation issues or know about the procedures and/or the legal issues. I do not consider other translators experts.

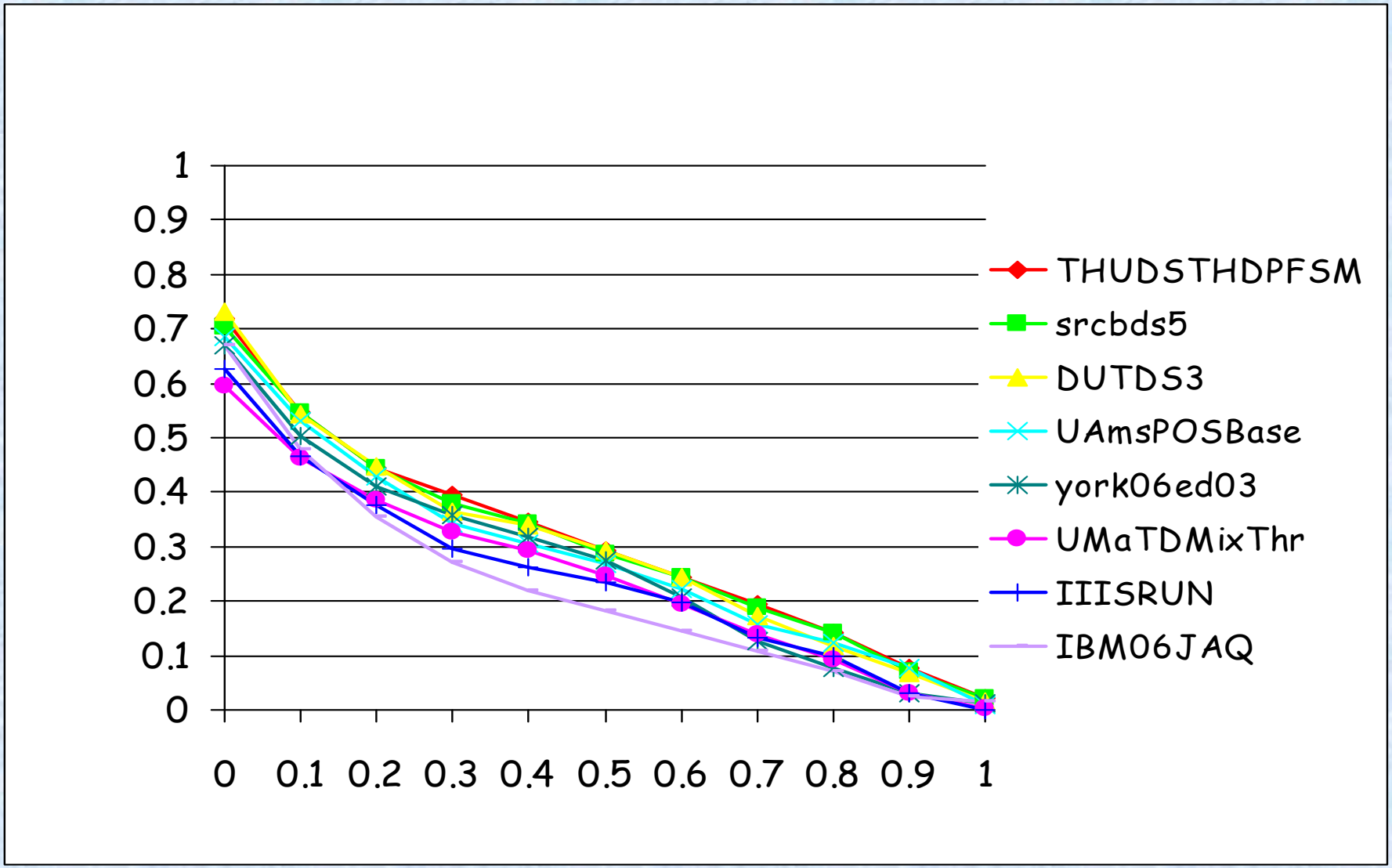
Search-for-Experts



Discussion Search

- Retrieve (email) documents containing arguments pro/con about a decision
  - discussion list portion of collection only
  - topics created at NIST based on categories produced from last year's track
  - messages judged not relevant, on-topic, containing argument
    - assessor marked whether argument was pro, con, both but this was not part of system's task

Discussion Email Search



Blog Track

- New track in 2006
  - explore information access in the blogosphere
- One shared task
  - find opinions about a given target
  - traditional ad hoc retrieval task evaluation, similar to discussion search in enterprise track
- Document set
  - set of blogs collected in Dec 2005-February 2006 & distributed by University of Glasgow
  - document is a permalink: blog post plus all its comments
  - ~3.2 million permalinks, some other info collected

Topics & Judgments

- 50 topics
  - created at NIST by backfitting queries from blog search engine logs
  - title was submitted query; NIST assessor created rest of topic to fit
- Judgments
  - judgments distinguished posts containing opinions from simply on-topic posts; also marked polarity
  - systems not required to give polarity, but did have to distinguish opinions from on-topic
  - assessors could skip document in pool if post likely offensive, but never exercised that option

Sample Topics

<title> ``March of the Penguins''

<description>
Provide opinion of the film documentary ``March of the Penguins''.

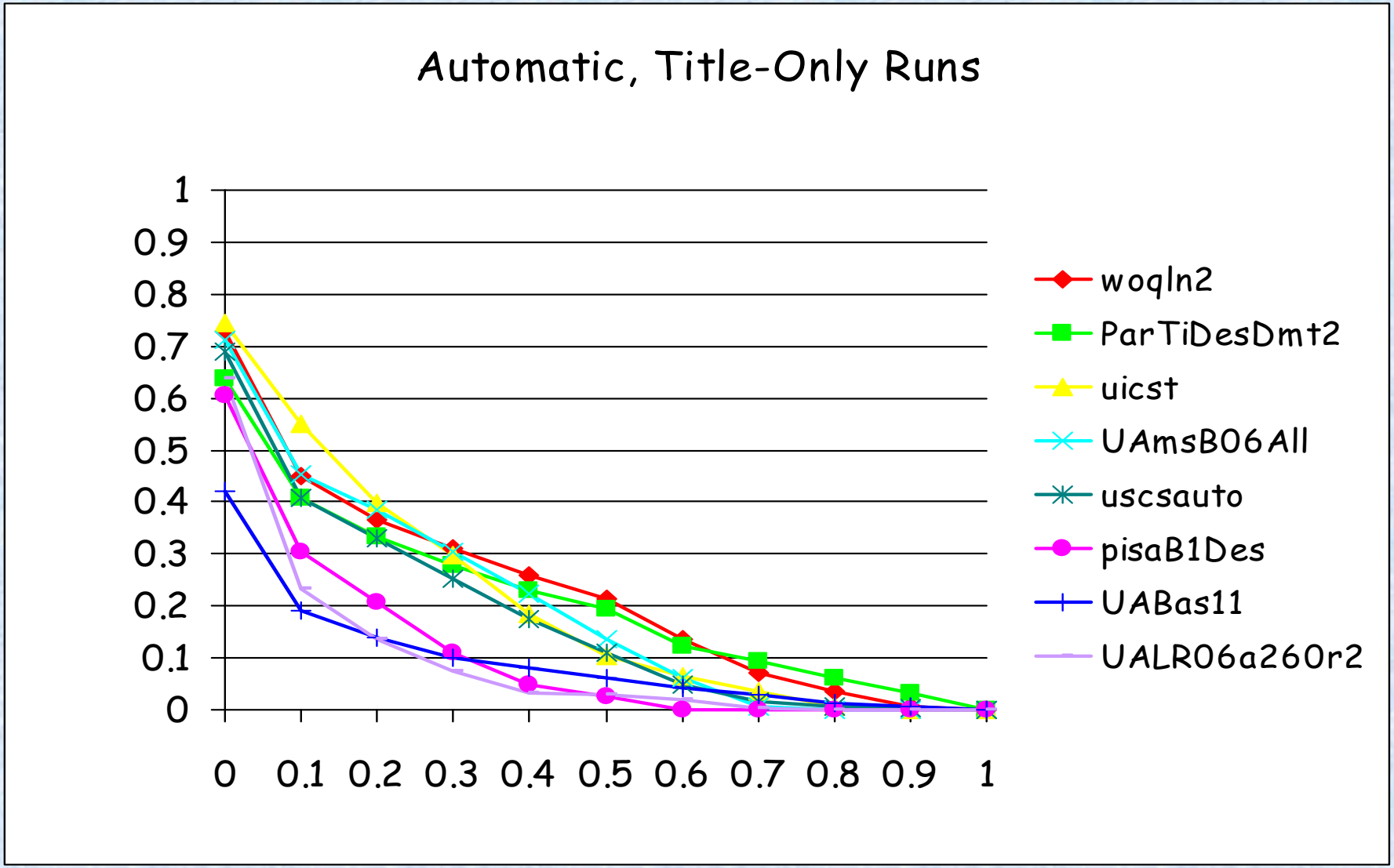
<narrative>
Relevant documents should include opinions concerning the film documentary ''March of the Penguins''. Articles or comments about penguins outside the context of this film documentary are not relevant.

<title> larry summers

<description>
Find opinions on Harvard President Larry Summers' comments on gender differences in aptitude for mathematics and science.

<narrative>
Statements of opinion on Summers' comments are relevant. Quotations of Summers without comment or references to Summers' statements without discussion of their content are not relevant. Opinions on innate gender differences without reference to Summers' statements are not relevant.

Opinion Results



Legal Track

- Goal: contribute to the understanding of ways in which automated methods can be used in the legal discovery of electronic records
- Timely track in that changes to the US Federal Rules of Civil Procedure regarding electronic records take effect Dec. 1

Legal Track Collection

- Document set
  - almost 7 million documents made public through the tobacco Master Settlement Agreement
  - snapshot of the University of California Library's Legacy Tobacco Document Library produced by IIT CDIP project [IIT CDIP Test Collection, version 1.0]
  - documents are XML records including metadata and a text field consisting of the output of OCR
  - wide variety of document types including scientific reports, memos, email, budgets…

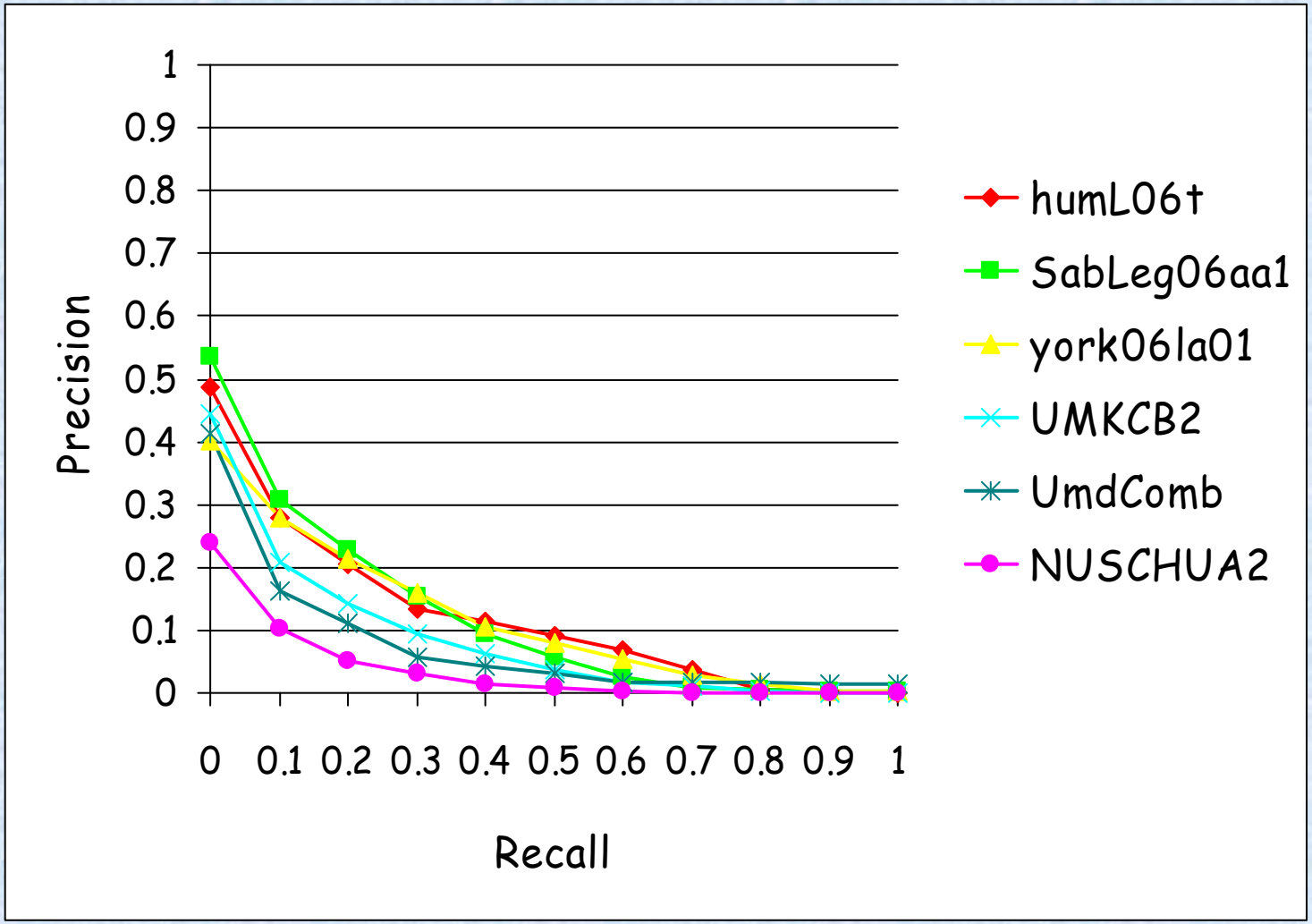
Legal Track Collection

- Topic set
  - modeled after actual legal discovery practice and created by lawyers
  - five (hypothetical) complaints with several requests to produce (topics) per complaint
  - also included a negotiated Boolean query statement that participants could use as they saw fit
  - 46 topics distributed; 39 used in scoring

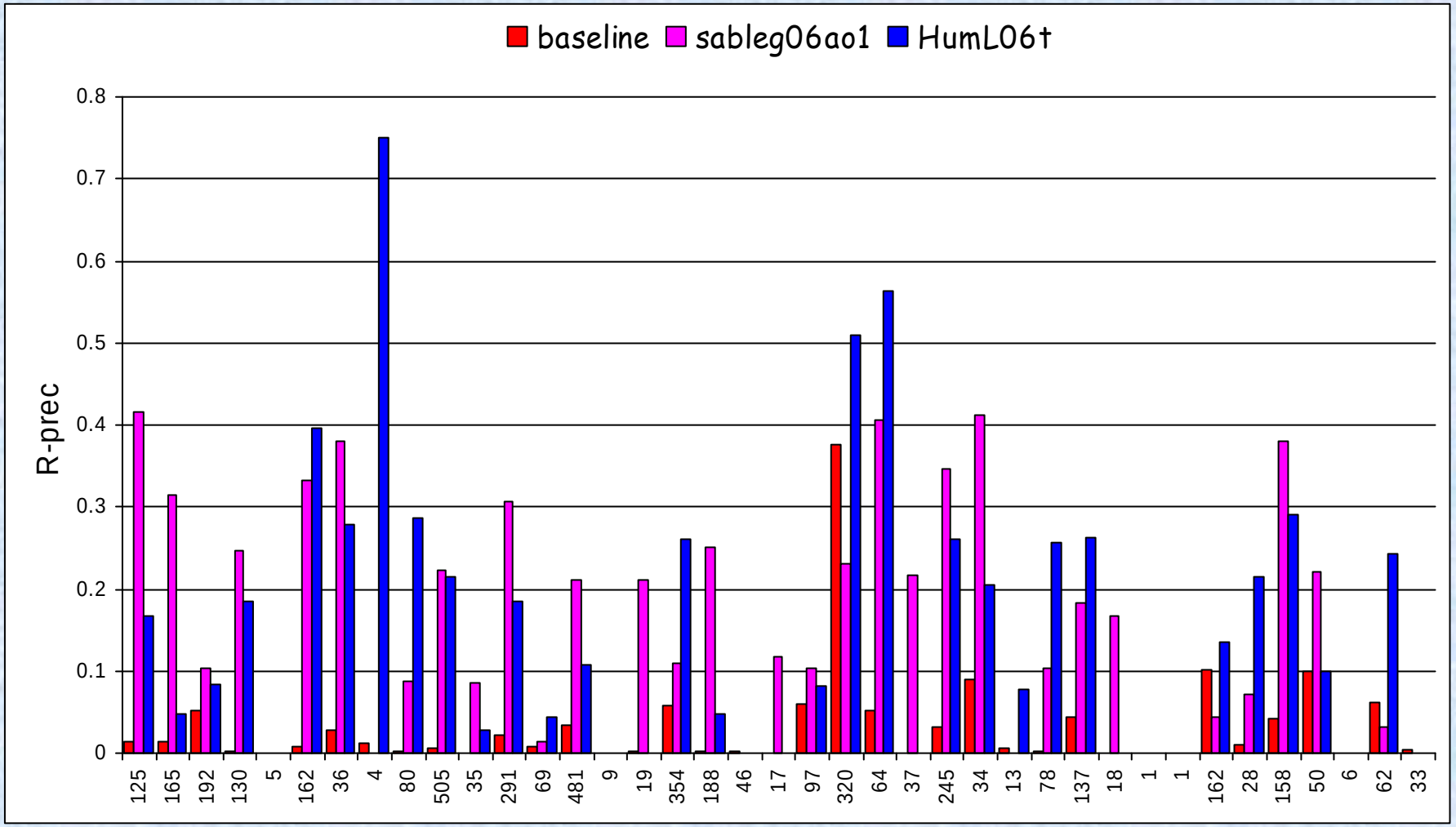
Legal Track Collection

- Relevance judgments
  - pools built from participant's submitted runs (λ=100 for one run per participant & λ=10 for remaining runs) …
  - plus set of documents retrieved by a manual run created by a document set expert who targeted unique relevant docs (approx. 100 docs/topic) …
  - plus a stratified sample of a baseline Boolean run (up to 200 documents/topic)
  - judgments made by law professionals

Ranked Retrieval Results



R-Precision Results



Text REtrieval Conference (TREC)

Spam Track

- Motivation:
  - assess quality of an email spam filter's actual usage
  - lay groundwork for other tasks with sensitive data
- How to get appropriate corpus?
  - true mail streams have privacy issues
  - simulated/cleansed mail streams introduce artifacts that affect filter performance
  - track solution: create software jig that applies given filter to given message stream and evaluates performance based on judgments
  - have participants send filters to data

Spam Tasks

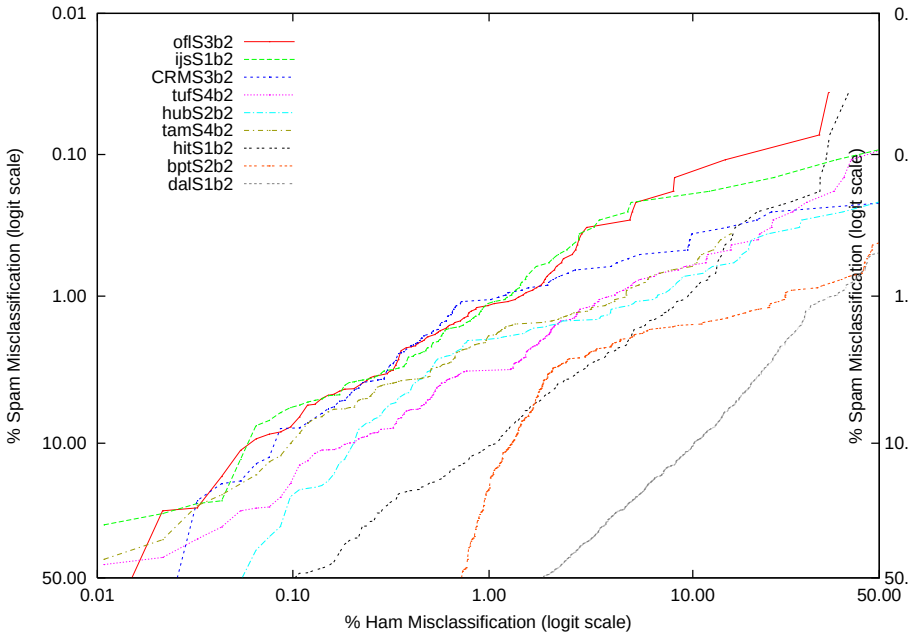
- 4 email streams [ham; spam; total]
  - public English [12,910; 24,912; 37,822]
  - public Chinese [21,766; 42,854; 64,620]
  - MrX2 (private) [9039; 40,135; 49,174]
  - SB2 (private) [9274; 2695; 11,969]
- 3 tasks
  - immediate feedback filtering
  - delayed feedback filtering
  - active learning

Evaluation

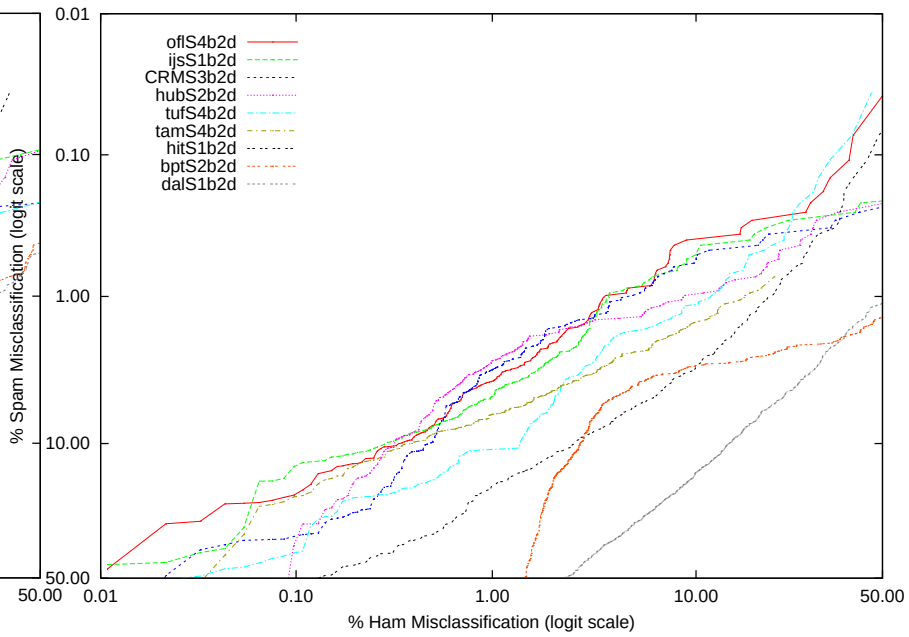
- Ham misclassification rate (hm%)
- Spam misclassification rate (sm%)
- ROC curve
  - assumes filter computes a "spamminess" score
  - use score to compute sm% as function of hm%
  - area under ROC curve is measure of filter effectiveness
  - use 1-area expressed as a % to reflect filter ineffectiveness (1-ROCA)%

Immediate vs. Delayed Feedback

SB2 Corpus

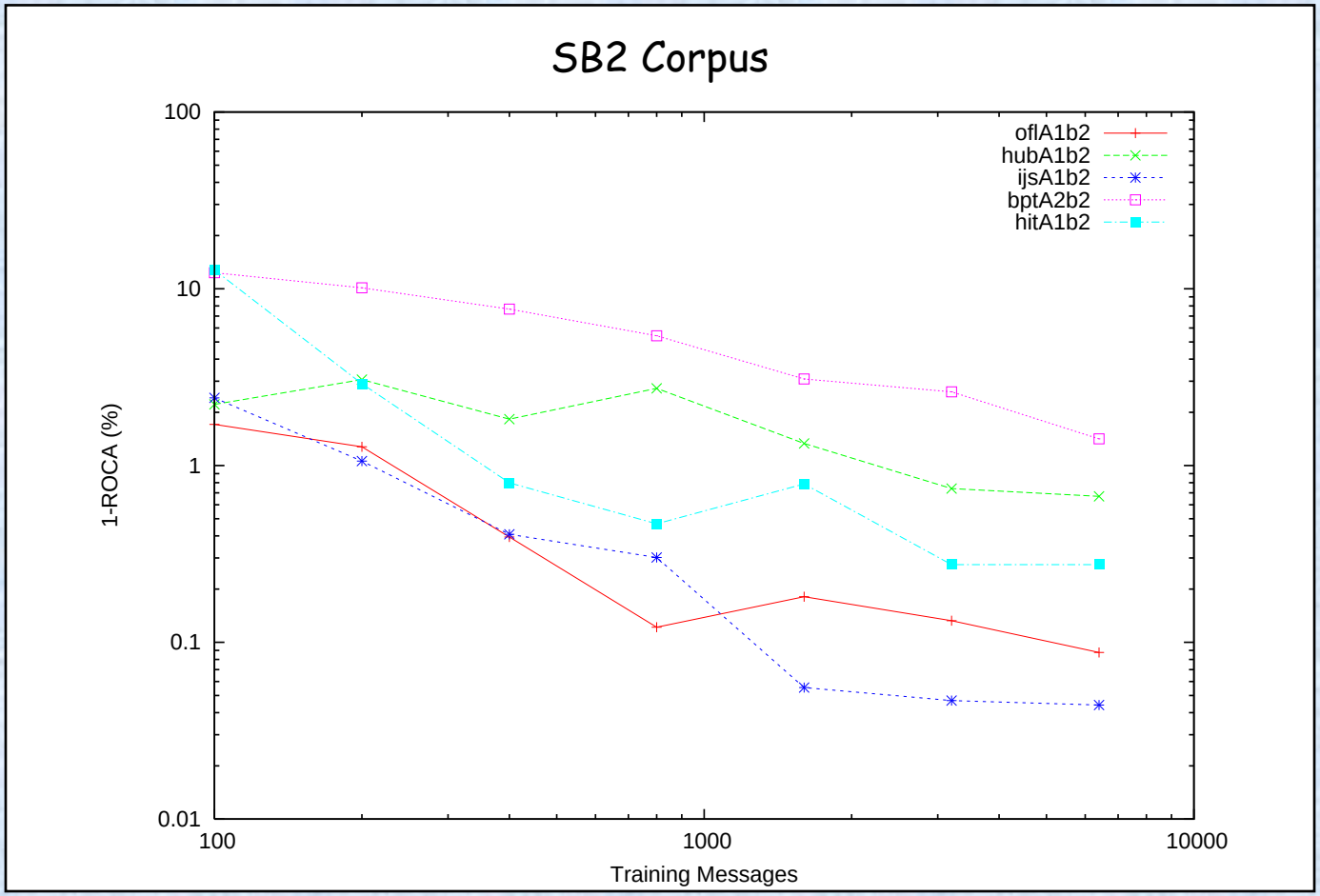


Immediate Feedback



Delayed Feedback

Active Learning



Future

- TREC expected to continue into 2007
- TREC 2007 tracks:
  - all but terabyte continuing
  - adding "million query" track
    - goal is to test hypothesis that a test collection built from very many very incompletely judged topics is a better tool than a traditional TREC collection
  - start planning for an eventual Desktop search track
    - goal is to test efficacy of search algorithms for the desktop
    - privacy/realism considerations severe barriers to traditional methodology

Track Planning Workshops

- Thursday, 4:00-5:30pm
  - Blog, LR B
  - Desktop, LR A
  - Genomics, Green
  - Legal, LR E
  - Terabyte, LR D
- Friday, 10:50am-12:20pm
  - Enterprise, LR B
  - Million Query, LR A
  - QA, Green
  - Spam, LR D