Robust Retrieval Track Overview

Ellen Voorhees

NIST

National Institute of Standards and Technology

Technology Administration, U.S. Department of Commerce

Robust Retrieval Track

- Motivations:
  - focus on poorly performing topics since average effectiveness usually masks huge variance
  - bring traditional ad hoc task back to TREC
- Task
  - 100 topics
    - 50 old topics from TRECs 6-8
    - 50 new tropics created by 2003 assessors
  - TREC 6-8 document collection: disks 4&5 (no CR)
  - standard trec_eval evaluation plus new measures

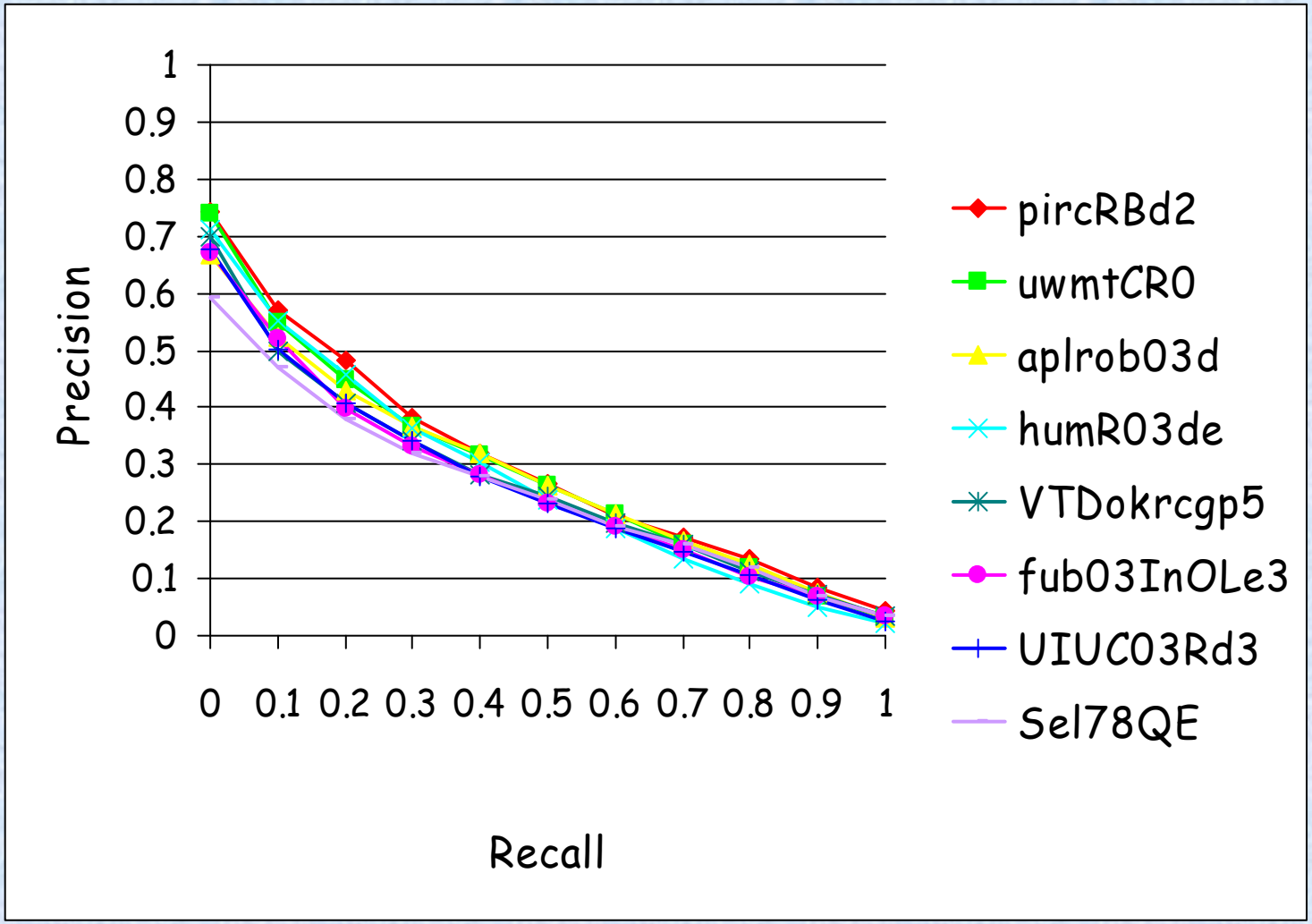
Robust Submissions

- 78 runs from 16 groups

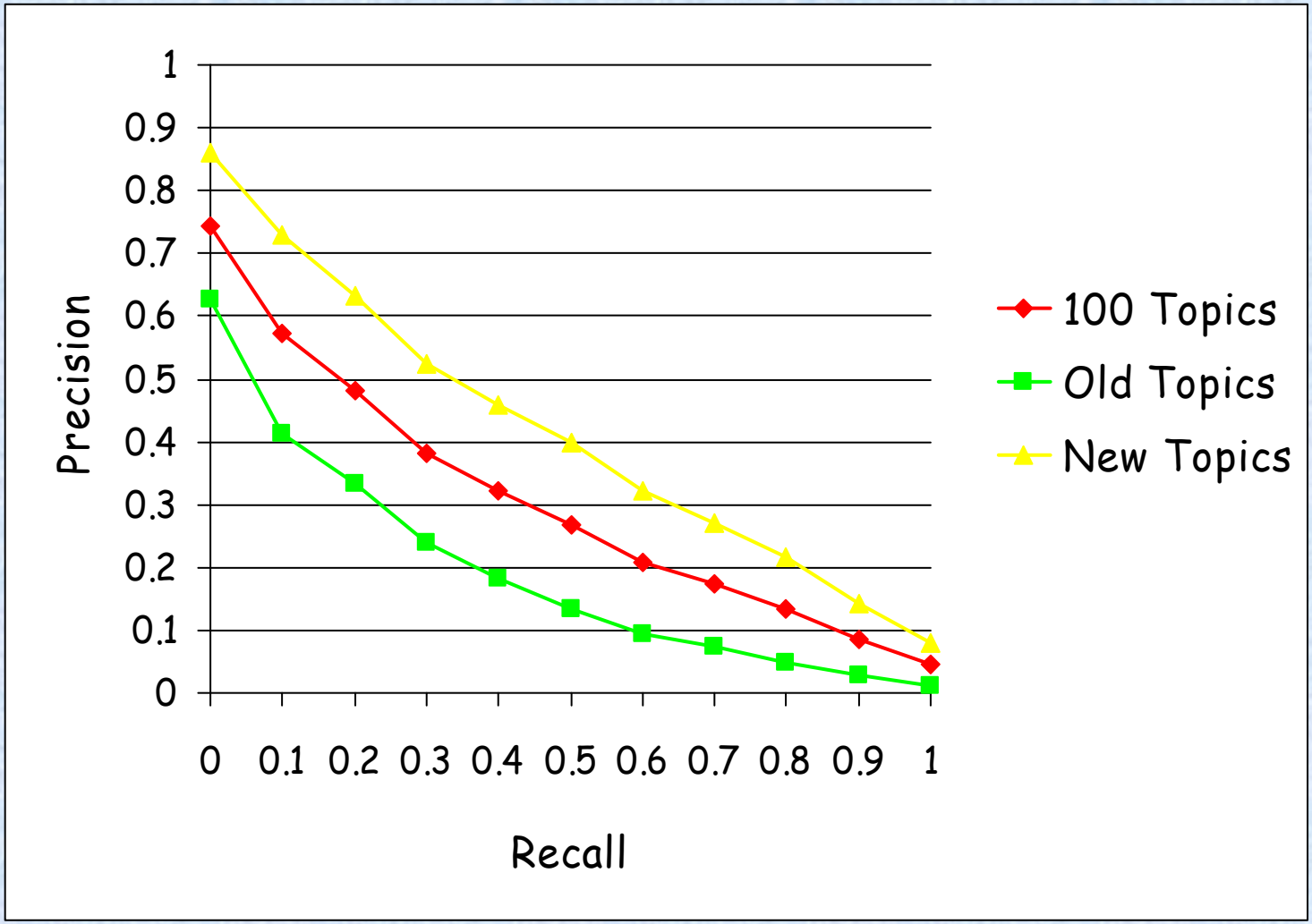
CAS-NLPR	Tsinghua U.
Fondazione Ugo Bordoni	U. of Amsterdam
Hummingbird	U. of Glasgow
Johns Hopkins/APL	U. of Illinois at Chicago
OcE technologies	U. of Illinois Urbana-Champaign
Queens College, CUNY	U. of Melbourne
Rutgers U.	U. of Waterloo
Sabir Research	Virginia Tech

- All runs automatic
- Description-only required run
  - topic length had significant effect

Best Description-Only Runs, Combined Topic Set



R-P Curves for Different Topic Sets



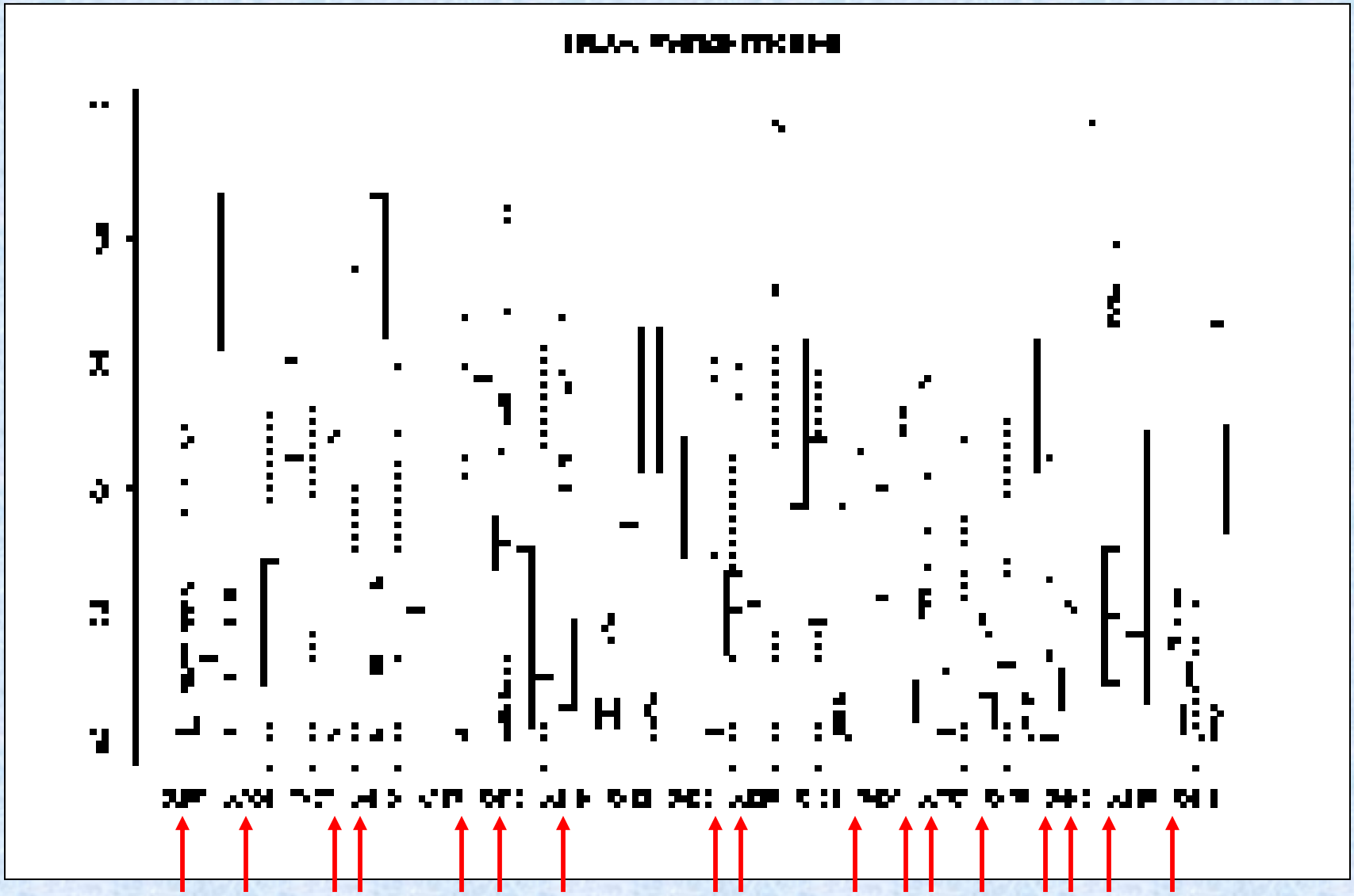
Retrieval Methods

- CUNY and Waterloo expanded using the web (and possibly other collections)
  - effective, even for poor performers
- QE based on target collection generally improved mean scores, but did not help poor performers
- Approaches for poor performers
  - predict when to expand
  - fuse results from multiple runs
  - reorder top ranked based on clustering of retrieved set

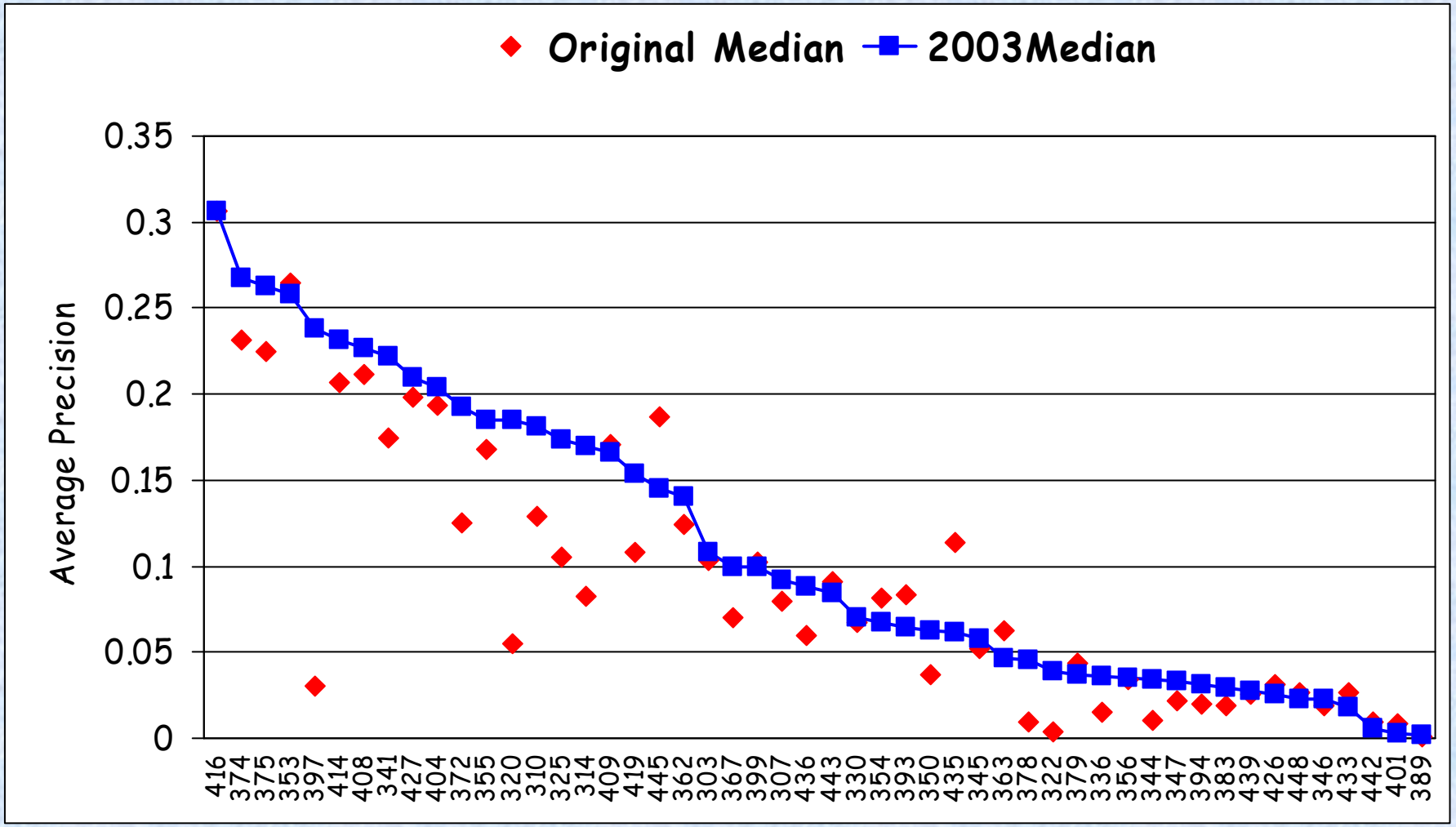
Old vs. New Topic Sets

- 50 old topics known to be difficult
  - median average precision score low with at least one high outlier in previous TREC
  - relevants: mean 88, min 5, max 361
  - systems may have trained on them
- 50 new topics intended as control group
  - created using standard topic development process
  - relevants: mean 33, min 4, max 115

Selecting Old Topics



Comparison of Median Scores
(50 Old Topics)



Measures for Robust Retrieval

- Percentage of topics with no relevant retrieved in top 10
  - direct, intuitive measure of behavior of interest
  - very coarse measure
- Area under MAP(X) vs. X curve
  - much more sensitive but far less intuitive measure
  - compute MAP over worst X topics & plot value as a function of X; use $X \leq \frac{1}{4}N$ when there are N topics total; calculate area underneath this curve
  - emphasizes the worst topics
  - different systems have different worst topics, so measure computed over different set per system

Old vs. New Topic Sets

	System Rankings (old/new)	τ
MAP	WXCVoDLAqBHIFErhJnimNjpGlkegfMdRUOTQKSPcbZaY qWoVXCrLnljIEBmiHNADFpGhMJfegdkUORTQKSPcZbaY	0.772
P(10)	WXoLqIFERQPHVrjGpTSJhiCNgnDBmAMOlkUdefKcZbaY oWXqVjrFnClBImLGENpJMHeRQPifAhOUgkDTSdKZcbaY	0.562
% no	DRQPTSWXoGkgjArpOKqMJLBHIEmUFhnldCViNefZbcaY WVXpqojGMJgRQPIFfdOTSrlEBAeKLNDCnmHhkUiZcbaY	0.427
area	qojWpBIADXkEHMCgdRrGJFVhLOTfQmnileUSKPNcZabY WVqXojBApfDeCdJMGrLlOmFNngIkURKEHTQihPSZcbaY	0.560

Large differences in relative performance for different topic sets:

- different amounts of training on old topics
- different abilities to handle difficult topics

Robust Measures

	Old Topics			New Topics			All Topics		
	P(10)	% no	area	P(10)	% no	area	P(10)	% no	area
MAP	0.560	0.171	0.558	0.753	0.334	0.588	0.592	0.180	0.584
P(10)		0.433	0.444		0.463	0.535		0.397	0.493
% no			0.393			0.518			0.457

Kendall τ scores for rankings produced by different measures

Large differences in relative performance for different measures

- % topics with no relevant unstable measure? or
- MAP very unaffected by poor performers? or
- ???

Conclusions

- Robust retrieval track provided
  - strong confirmation that traditional average effectiveness measures do not reflect poorly performing topics
  - evidence that difficult topics are still difficult
- Open questions
  - What are the implications of the differences in topic sets for collection building?
  - Are the new measures
    - stable?
    - meaningful?
    - useful?