

The TREC Question- Answering Track

Donna Harman
Retrieval Group
Information Access Division
National Institute of Standards and
Technology

IR Problems and TREC Tasks

Answers, not documents

Web searching

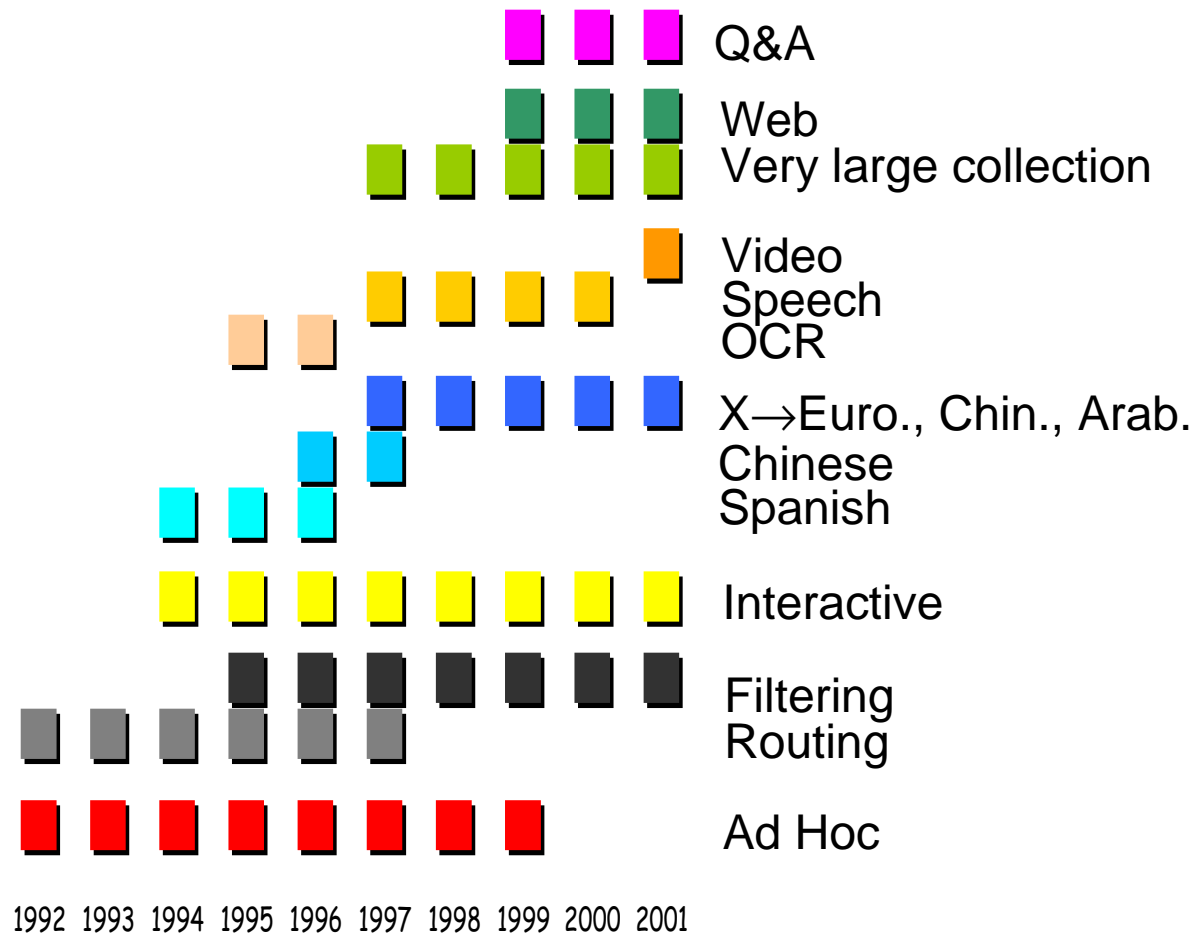
Beyond text

Beyond just English

Human-in-the-loop

Streamed text

Static text



Question Answering Track

- Goal: encourage research into systems that return answers, rather than document lists
 - 3 subtasks in 2001
 - main
 - retrieve 50-byte snippet of text containing answer
 - list
 - assemble a set of instances as the answer to a question
 - context
 - track discourse objects through a series of questions and answers



Alicante University	Korea University	Syracuse U.
Chinese Acad. Sci.	LCC	Tilberg U.
CL Research	LIMSI	U. of Alberta
Conexor Oy	Microsoft	U. of Amsterdam
EC Wise, Inc.	MITRE	U. of Illinois, U-C
Fudan University	National Taiwan U.	U. of Iowa
Harbin Inst.	NTT Comm. Labs	U. de Montreal
IBM Research (1)	Oracle	U. of Pennsylvania
IBM Research (2)	National Taiwan U.	U. of Pisa
InsightSoft-M	Oracle	USC-ISI
ITC-irst	Pohang U.	U. of Waterloo
KAIST	Queens College	U. of York
KCS	Sun Microsystems	

Question Type

- Open domain, closed-class questions

How far is it from Denver to Aspen?

What county is Modesto, California in?

What is an atom?

When did Hawaii become a state?

- Answers less than 50 characters
- Answers generally named entities or noun phrases

Document Collection

- Answers to be found in a large corpus of news articles
 - newspaper and newswire documents on TREC disk 1-5
 - approximately 3 gb of text
 - approximately 979,000 articles

Question Source

- Progression into more “real” questions
 - TREC-8: created for track
 - TREC-9: ideas mined from logs
 - TREC 2001: questions directly from filtered logs
 - MSNSearch & AskJeeves logs restricted to queries with question words & modals
 - NIST further winnowed by hand
 - NIST assessors verified answers

TREC 2001 QA Main Task

- Final test set of 500 questions
 - NIST fixed punctuation, spelling
 - 8 questions eventually removed from test set due to various problems
 - 49 had no known answer in collection
 - assessor didn't find answer and no system returned a correct (supported) answer
 - NIL was correct response for these

Response Unit

- Systems return [docid, answer-string] pairs
 - answer-string required to be ≤ 50 bytes
 - answer-string must contain answer
 - document docid supports answer
- Guidelines
 - all runs completely automatic
 - no changes once questions received
 - submit results within 1 week

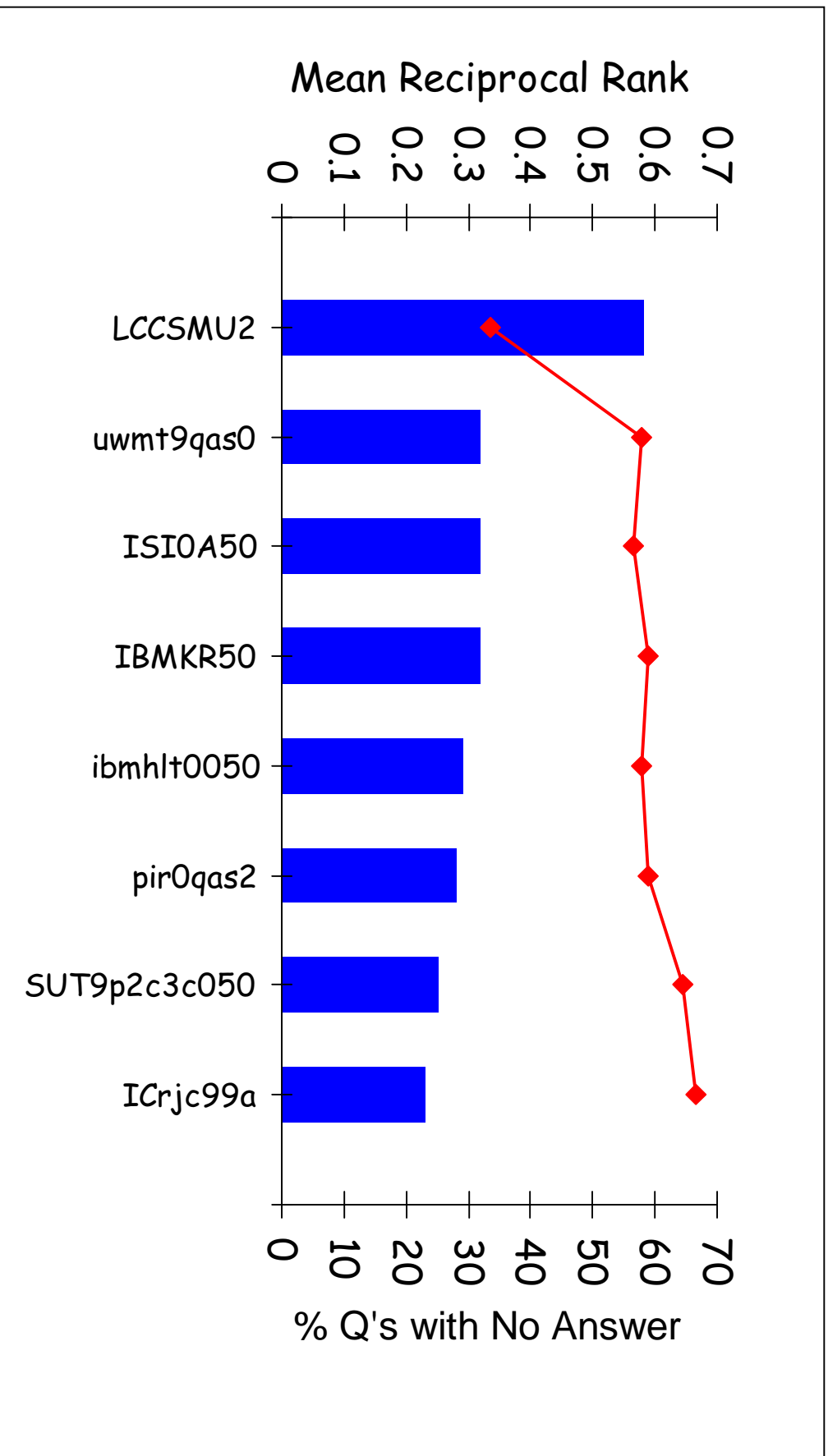
Human Judgments

- NIST assessors judge each answer string for correctness
 - 3-valued judgments:
correct, unsupported, incorrect
 - answer strings must be responsive
 - appropriate units
 - no answer stuffing
 - match assessor's interpretation of question
 - e.g., location of Taj Mahal

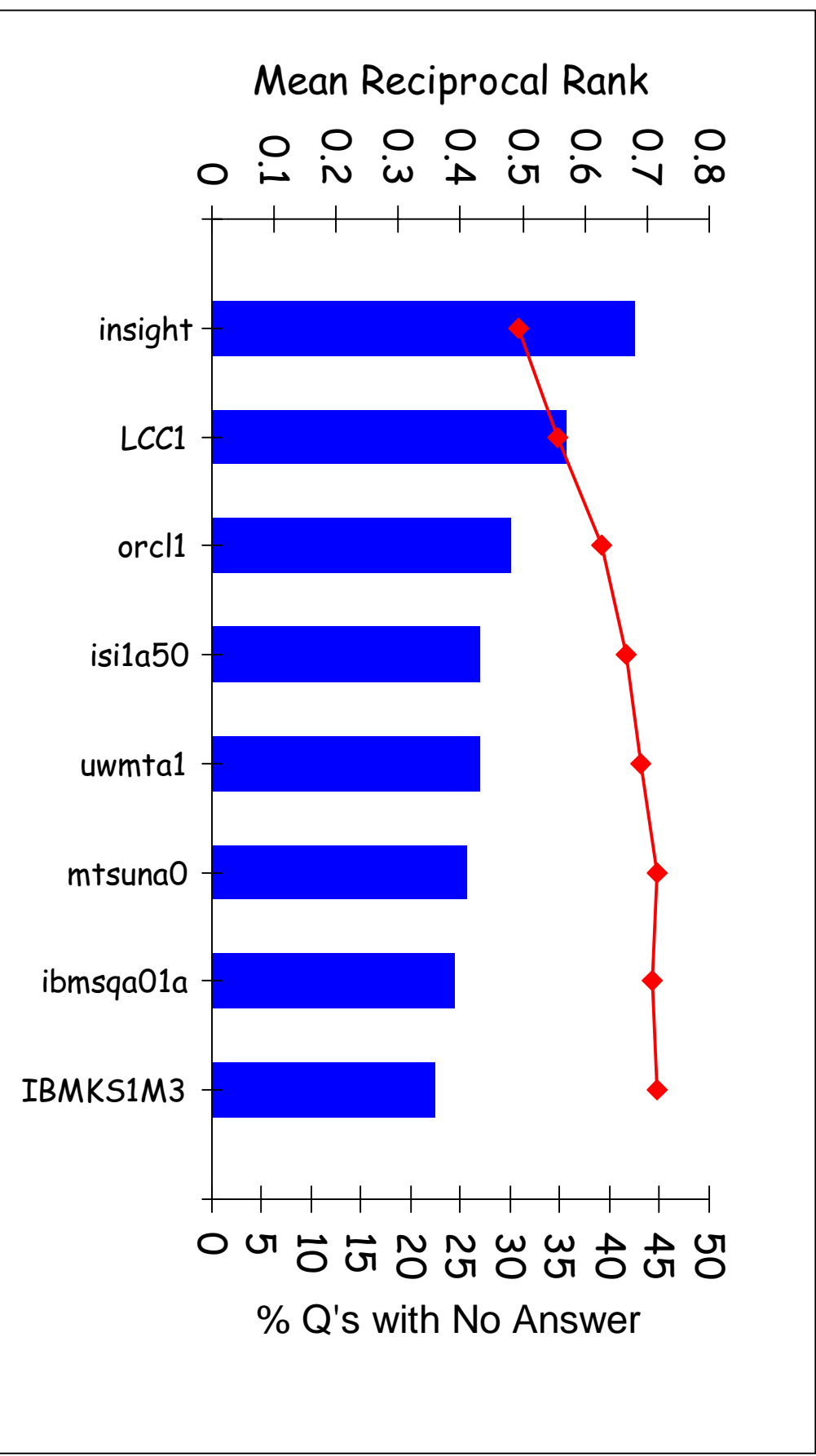
Assessor Opinions Differ

- TREC QA evaluations based on assumption that assessor opinions *WILL* differ
 - forcing agreement among TREC assessors doesn't address problem because deployed systems will have to deal with these variations and it is critical to model this environment
 - comparative evaluation is stable because of large numbers of questions, but only comparative evaluation is valid

TREC 2000 QA Results



TREC 2001 Main Task Results



Main QA 2001 Approaches

- Standard processing remains:
 - determine answer type from question form
 - retrieve small portion of documents
 - find correct answer type in document piece
- Majority of systems using a lexicon
 - usually WordNet
 - used to verify that candidate is of correct type
 - sometimes used for expansion

Differences in Approaches

- Debate over best type of answer type categorization
 - few broad classes vs. many specialized
 - trade-off between accuracy & coverage
 - some systems using hierarchy to exploit this trade-off
- Debate between data driven vs. deep understanding
 - machine learning in all steps
 - use of web to validate answer

Detecting No Answer

- Apparently, a difficult problem

- only 5 runs had accuracy > .25

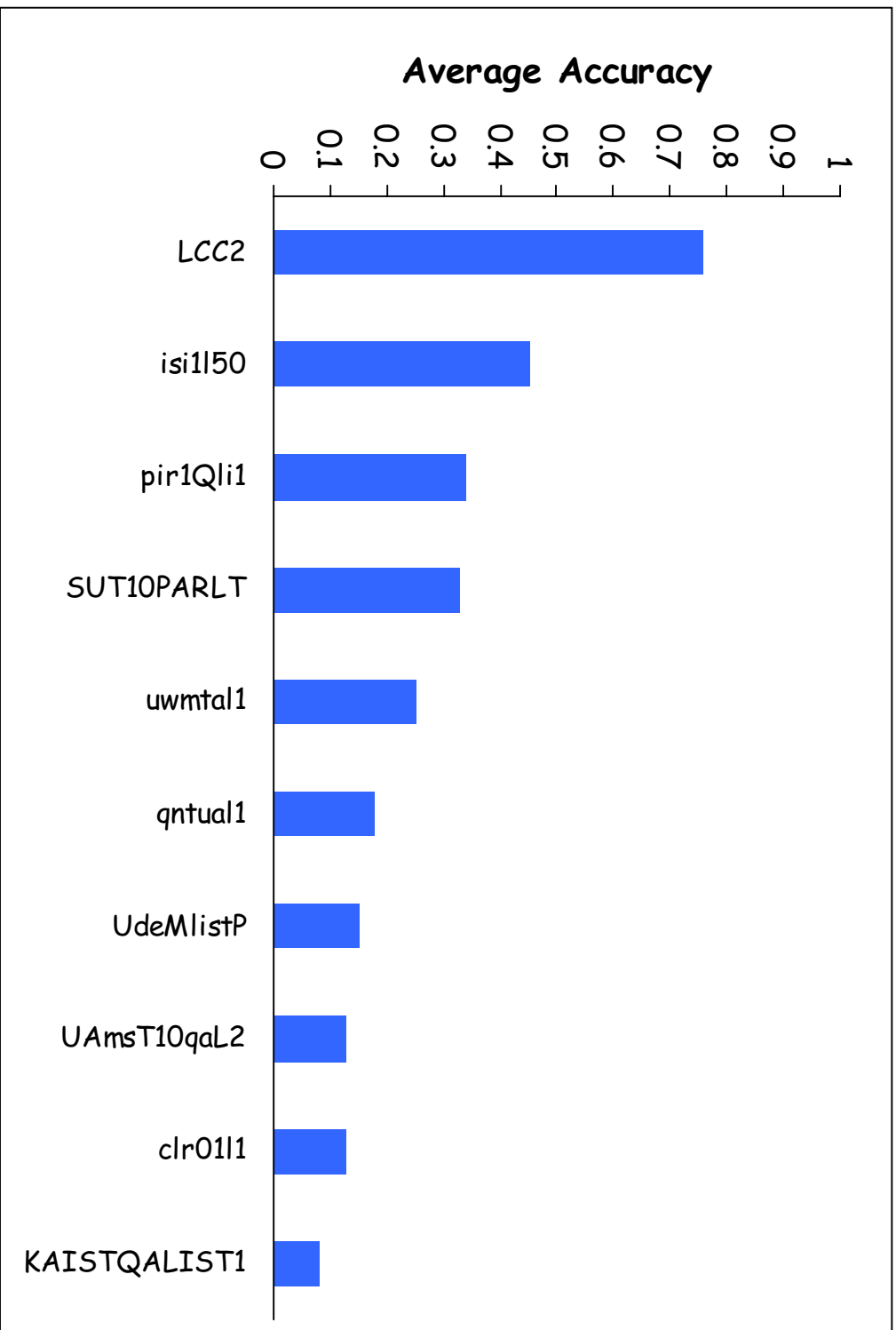
LCC1	.76	31/41
orcl1	.43	35/82
insight	.37	38/120
ICTQA10a	.29	10/35
ICTQA10b	.27	15/55

- returning NIL for all questions gives accuracy of .10

QA List Task

- Instance-finding task
 - 25 questions that specified a target number of instances to retrieve
 - *Name 4 U.S. cities that have a "Shubert" theater*
 - *What are 9 novels written by John Updike?*
 - response is an unordered set of the target number of instances
 - an instance is a single [docid, answer-string] pair
 - questions constructed by NIST assessors
 - target chosen such that collection had at least that number of instances but > 1 doc required
 - single document may have > 1 instance

QA List Results



What was learned from this focused evaluation?

- 38+ groups tried different methods, and in particular addressed different aspects of the problem; high potential for technology transfer
- Very large set of questions now available for analysis and training

Getting beyond short answers

- Large spectrum of research and evaluation issues
- New ARDA AQUAINT program
 - 6 year program in 3 phases
 - 16 contractors in 1st phase, kickoff meeting was in December 2001
 - Goal of addressing more complex questions plus issues of multilingual, multimedia

Evaluation issues

- More complex questions require more complex evaluation
- Examples:
 - Definitions, such as "who is X?"
 - Cause and effect, relationship chains
 - Narratives as answers
 - Questions in context

Future plans—stay tuned

- TREC 2002 will require exact answers, not just snippets of text
- 5 or 6 pilot evaluations of more complex question types for AQUAINT
- Pilot evaluation of interactive QA dialog in AQUAINT
- AQUAINT also involves several groups working with knowledge bases

trec.nist.gov

Text REtrieval Conference (TREC)

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

[Overview](#)

[Frequently Asked
Questions](#)

[Publications](#)

[Data](#)

[Information
for Active
Participants](#)



[Contact
Information](#)

[Past TREC
Results](#)

NIST

National Institute of Standards and Technology