Overview of the TREC 2002 Question Answering Track

Ellen Voorhees

NIST

National Institute of Standards and Technology

Technology Administration, U.S. Department of Commerce

Question Answering Track

- Goal: encourage research into systems that return answers, rather than document lists

- TREC 2002 is fourth year
  - as before, restricted to factoid questions with document for support
  - this year required exact answer, not text snippet

TREC 2002 QA Track

- Main task
  - return exactly one response for each of 500 questions
  - response is either [doc, string] pair or NIL
  - rank questions by confidence in answer
- List task
  - target number of instances given in question
  - assemble an unordered set of instances where an instance is a [doc, string] pair

QA Track Participation

Alicante University
BBN Technologies
CMU (JAVELIN)
Chinese Acad. of Sciences
CL Research
Columbia U.
Fudan University
IBM (Ittycheriah)
IBM (Prager)
InsightSoft-M
ITC-irst

Language Computer Corp.
LIMSI
MIT
MITRE
Nat'l U. Singapore (Lee)
Nat'l U. Singapore (PRIS)
NTT Commun. Science Labs
Pohang U. of Sci. & Tech.
Syracuse University
Tokyo U. of Science
Universite d'Angers
Universite de Montreal

University of Amsterdam
University of Avignon
U. Illinois, U-C
University of Iowa
University of Limerick
University of Michigan
University of Pisa
University of Sheffield
U. So. California, ISI
University of Waterloo
University of York

34 groups:

- 66 main task runs
- 9 list task runs from 5 groups

Data

- New AQUAINT document set
  - articles from NY Times newswire (1998-2000), AP newswire (1998-2000), and Xinhua News Agency (1996-2000)
  - approximately 3 gb of text
  - approximately 1,033,000 articles
- Questions taken from MSNSearch and AskJeeves logs
  - no definition questions
  - some spelling/grammatical errors remain
  - 46 questions with no known answer in docs

Motivation for Exact Answers

What river in the US is known as the Big Muddy?

- the Mississippi
- Known as Big Muddy, the Mississippi is the longest
- as Big Muddy , the Mississippi is the longest
- messed with . Known as Big Muddy , the Mississip
- Mississippi is the longest river in the US
- the Mississippi is the longest river in the US,
- the Mississippi is the longest river(Mississippi)
- has brought the Mississippi to ist lowest
- ipes.In Life on the Mississippi,Mark Twain wrote t
- Southeast;Mississippi;Mark Twain;officials began
- Known; Mississippi; US,; Minnesota; Gulf Mexico
- Mud Island,;Mississippi;”The;-- history,;Memphis

Motivation for Exact Answers

- Text snippets masking important differences among systems
- Pinpointing precise extent of answer important to driving technology
  - not a statement that deployed systems should return only exact answers
  - exact answers may be important as component in larger language systems

Exact Answers

- Human assessors judged responses
  - Wrong: string does not contain a correct answer or answer is unresponsive
  - Not Supported: string contains a correct answer, but doc does not support that answer
  - Not Exact: string contains correct answer and doc supports it, but string contains too much (or too little) info
  - Right: string is exactly a correct answer that is supported by the doc

Exact Answer Guidelines

- most minimal response possible not the only exact answer
  - e.g., accept "Mississippi river " for *What is the longest river in the United States?*
- ungrammatical responses not exact
  - e.g., "in Mississippi" vs. "Mississippi in"
- justification is not exact
  - e.g., "At 2,348 miles the Mississippi river is the longest US river" is inexact

Distribution of Judgments

- 15,948 judgments across all questions

12,639	79.3%	Wrong
505	3.2%	Unsupported
442	2.8%	ineXact
2,362	14.8%	Right

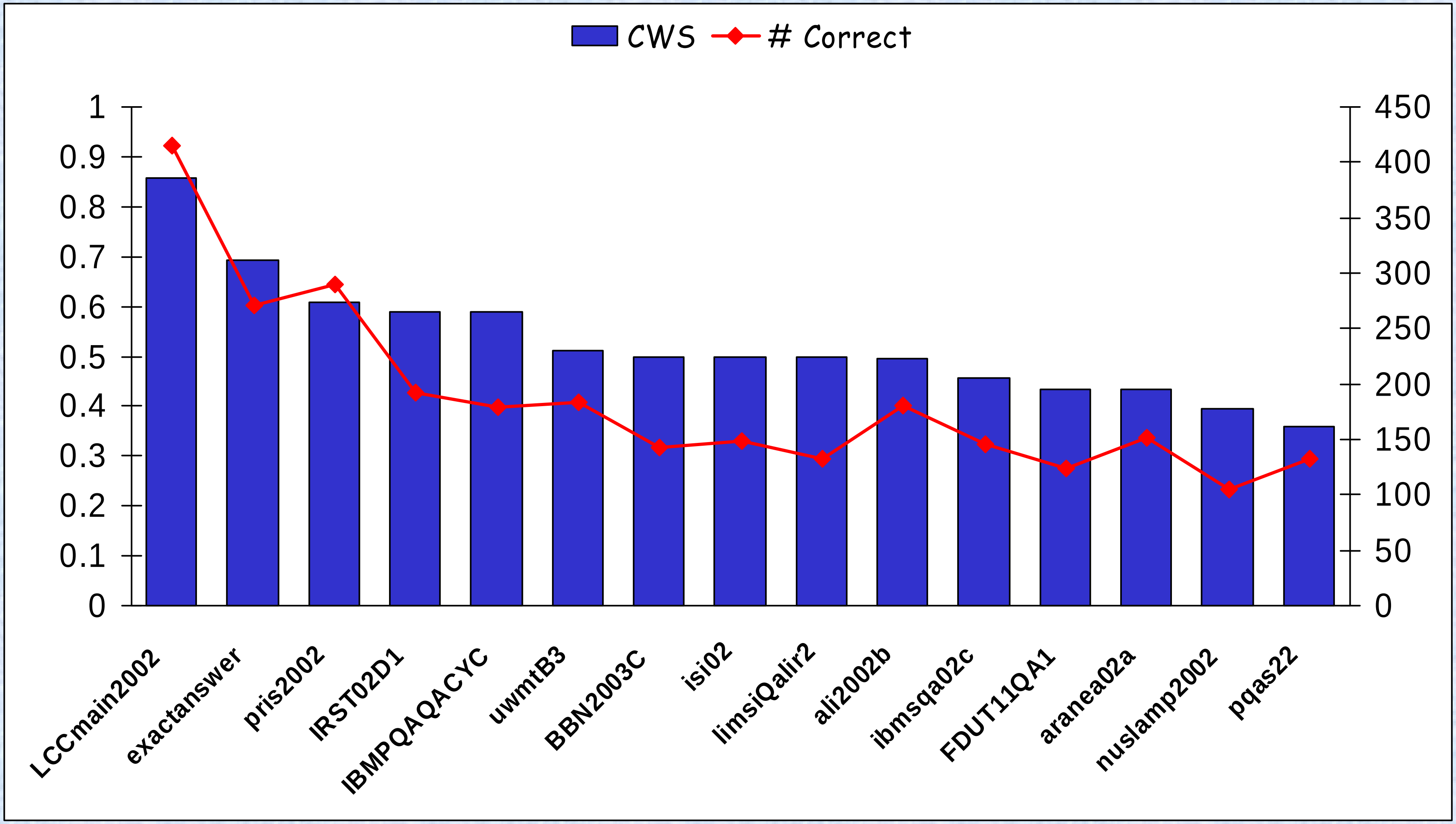
- In general, systems can find extent of answer if they can find it at all
  - distribution skewed across systems
  - attempt to get exact answer sometimes caused units to be lost (so marked wrong)

Confidence-weighted Scoring

- Focus on getting systems to know when they have found a good answer
  - questions ranked by confidence in answer
  - compute score based on ranking

$$\frac{\sum_{i=1}^{N} \text{number right to rank } i/i}{N}$$

Main Task Results



Main Themes

- Many systems now using specific data sources for expected question types
  - name lists
  - gazetteers
- Web used by most systems, but in different ways
  - primary source of answer that is then mapped to corpus
  - one of several sources whose results are fused
  - place to validate answer found in corpus

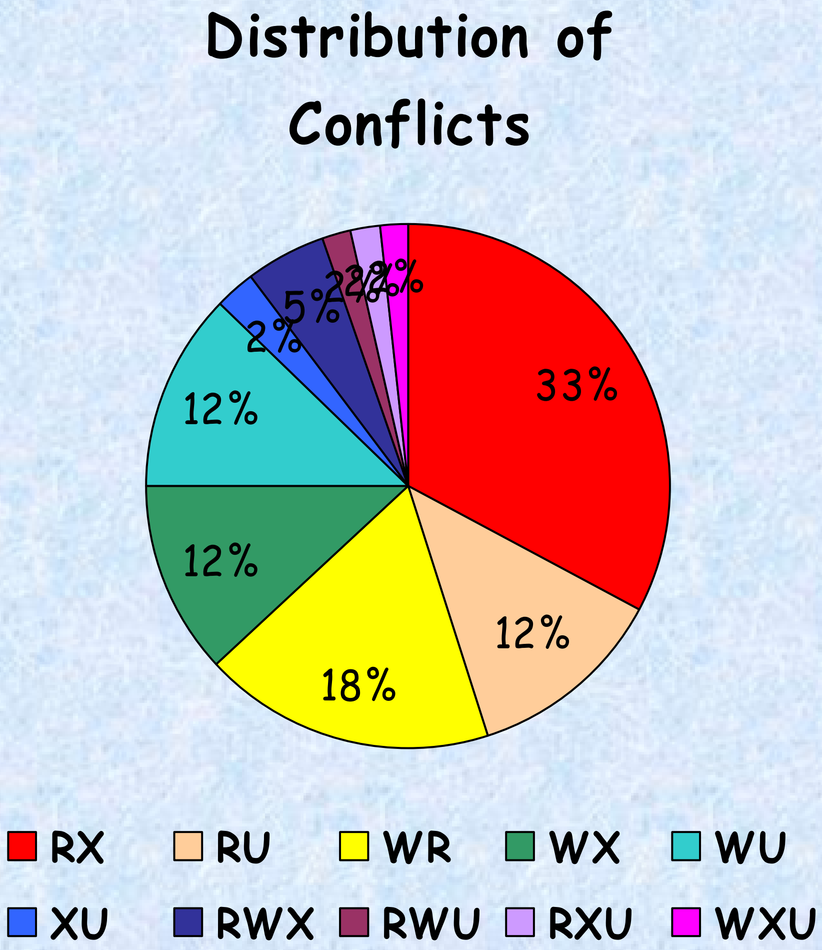
Confidence Ranking

- Different approaches
  - most groups used the type of question as a factor
  - some systems that use scoring techniques to rank candidate answers also used score for ranking questions
  - few groups used training set to learn good feature set and corresponding weights, then applied classifier to test set
  - many groups ranked NIL questions last

Quality of the Evaluation

- Assessors opinions differ but evaluation is stable when using text snippets and MRR metric. Now?
  - exact answers
  - single response per question
  - confidence-weighted score
- Repeat stability study using multiple independent assessments
  - each question judged by 3 assessors
  - official evaluation based on adjudicated judgments

Assessors Continue to Disagree



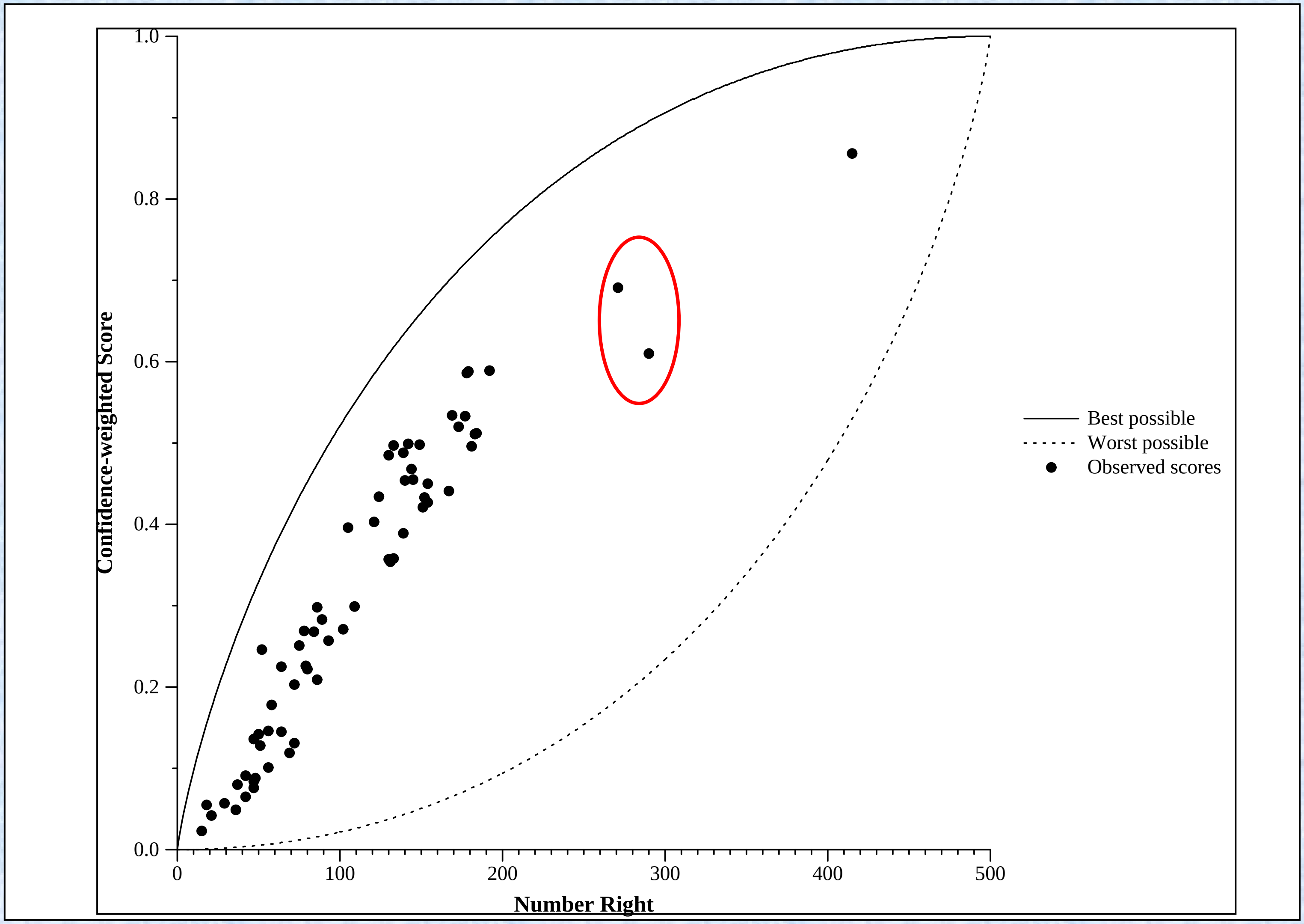
- 50% of judgments where at least one judgment was not W had disagreements
- Of those, 33% involved disagreements between Right and ineXact
  - well-known granularity issue now reflected here
- For dates and quantities, disagreement among Wrong and ineXact

Comparative Results Still Stable

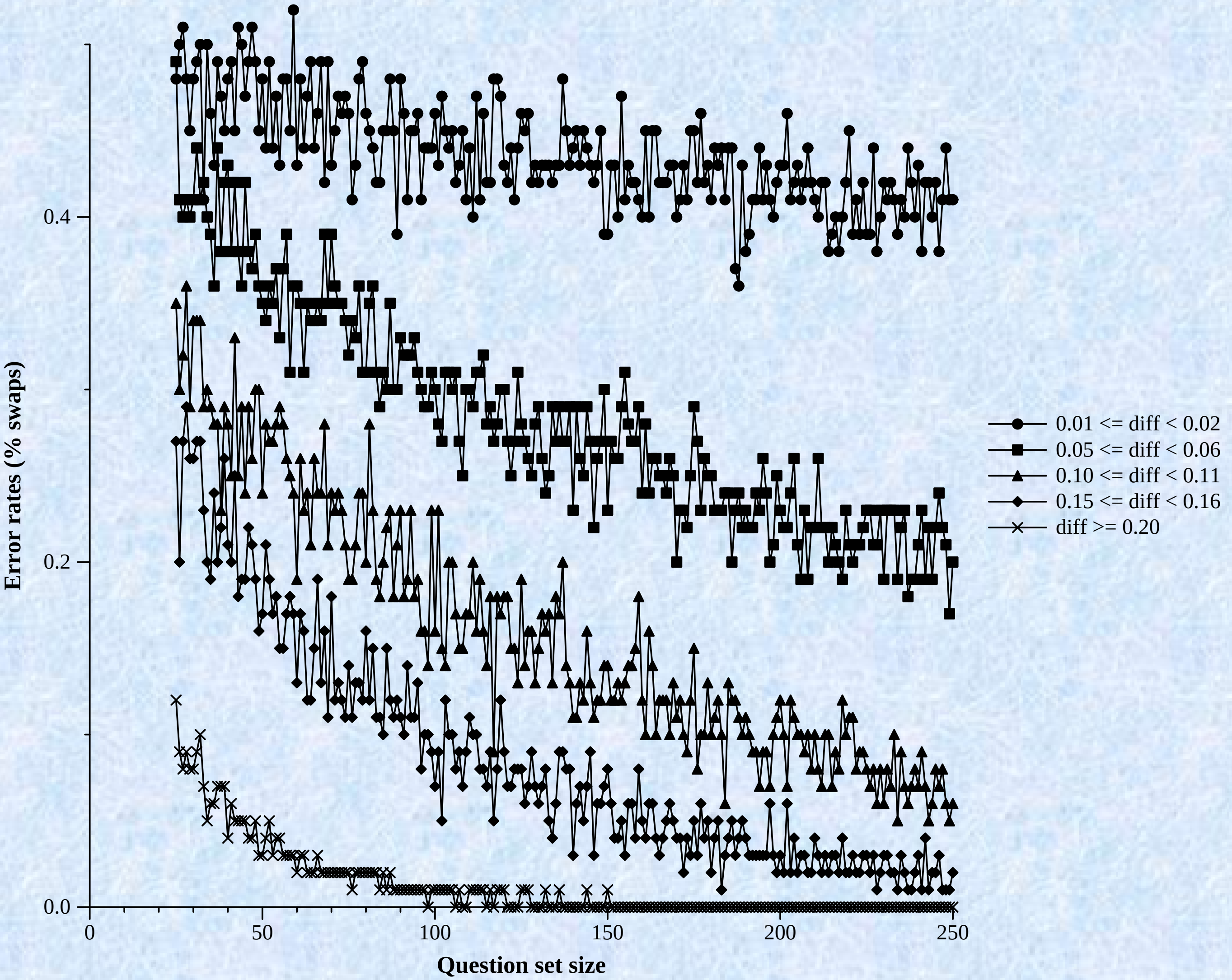
		Adj	1	2
Confidence weighted score	1	0.954		
	2	0.941	0.920	
	3	0.944	0.917	0.906
Number correct	1	0.958		
	2	0.949	0.933	
	3	0.960	0.944	0.926

- Kendall τ scores between system rankings > 0.9
- Scores for rankings using adjudicated judgments > 0.94
- Number correct measure more stable than confidence-weighted score

CWS Emphasizes Ranking



Inherent Stability of CWS



Summary

- Major changes in TREC 2002
  - exact answers
    - working definition of exact answer ok
    - in general, systems can detect answer extent
  - confidence ranking
    - CWS puts large emphasis on proper ranking
    - evaluation results stable with large enough question set