Overview of the TREC 2003 Question Answering Track

Ellen Voorhees

NIST

National Institute of Standards and Technology

Technology Administration, U.S. Department of Commerce

Tasks

- Passages task
  - return single document extract per factoid question
- Main task
  - return single exact answer per factoid
  - return set of factoid answers for lists
  - return description of target for definitions

Data

- 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
  - factoids and definition questions taken directly from MSNSearch and AOL logs
  - list questions created by assessors

Passages Task

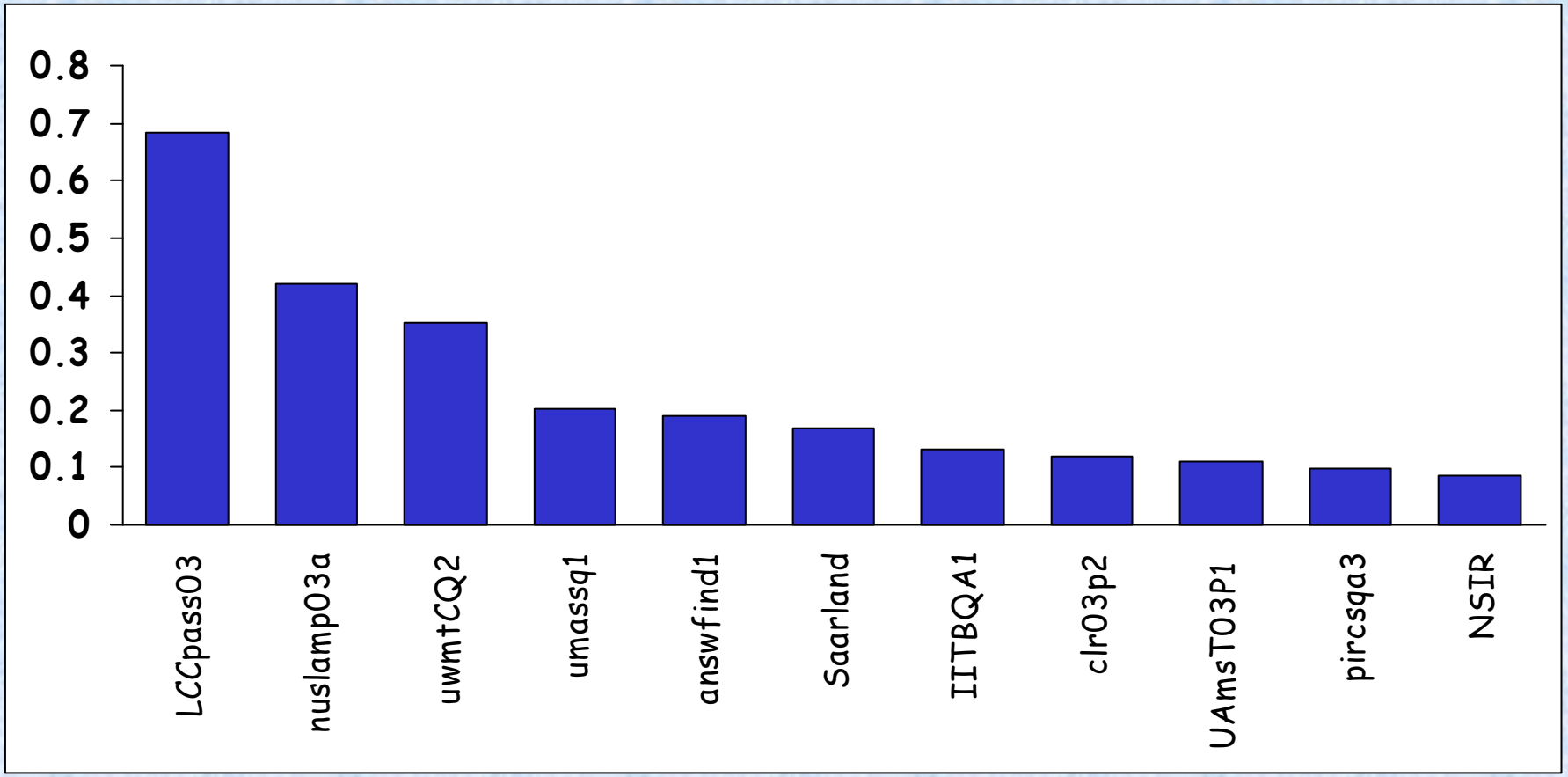
- Return a document extract as an answer to a factoid question
  - extract restricted to ≤250 characters
  - document extracts to avoid answer stuffing
- Answers
  - judged correct, unsupported, or wrong
  - independently judged by two assessors; differences adjudicated for final evaluation
- Score is accuracy, percentage of questions answered correctly

Motivation for Extracts

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

Passages Task Results

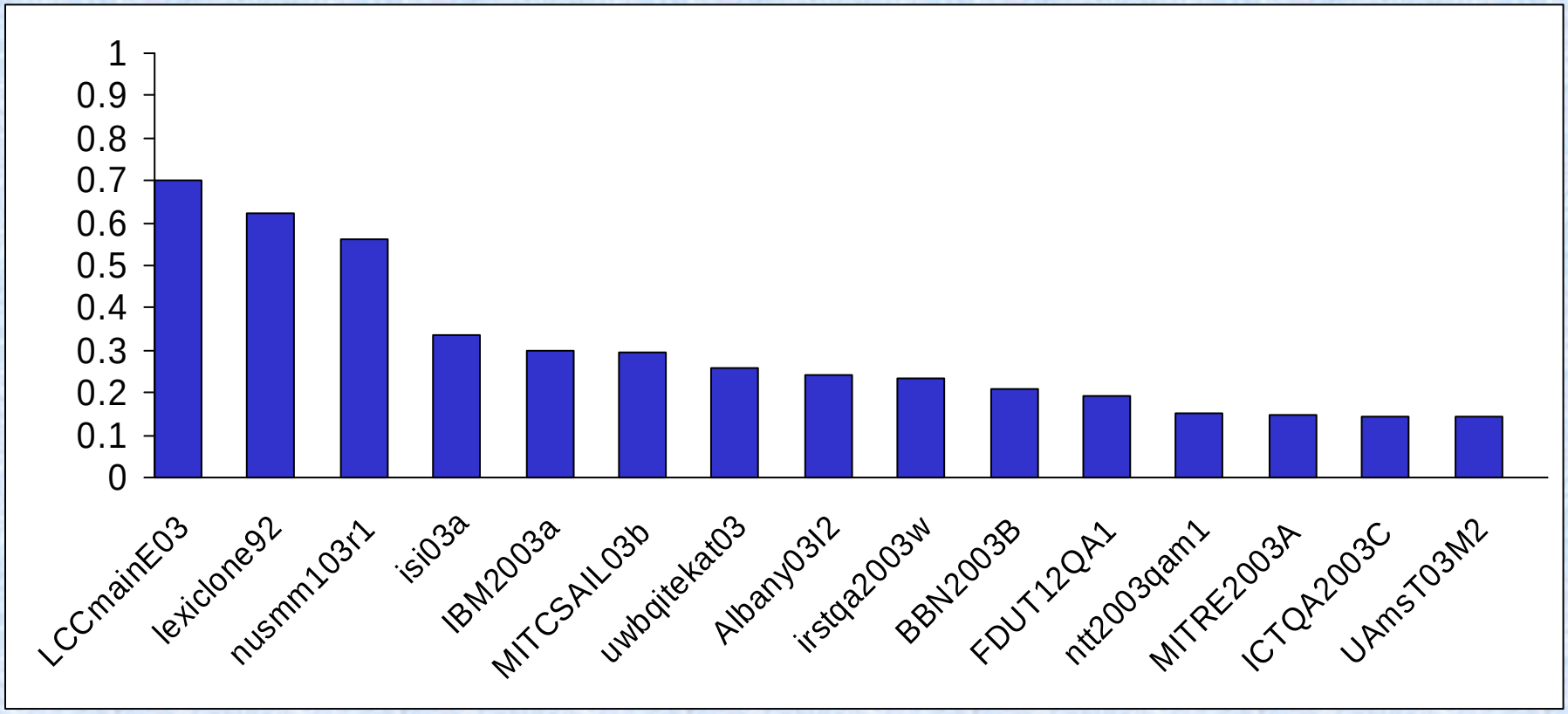


Accuracy for best passages task run per group

Factoid Component of Main Task

- Same as passages task except exact answer required
  - responses could also be judged "inexact"
- 3 groups did passages and main tasks
  - factoid accuracy (insignificantly) greater in main task than passages task for 2/3
- Accuracy $\frac{1}{2}$ of final main task score

Main Task Factoids Results



Accuracy of best run for top 15 groups for factoid component

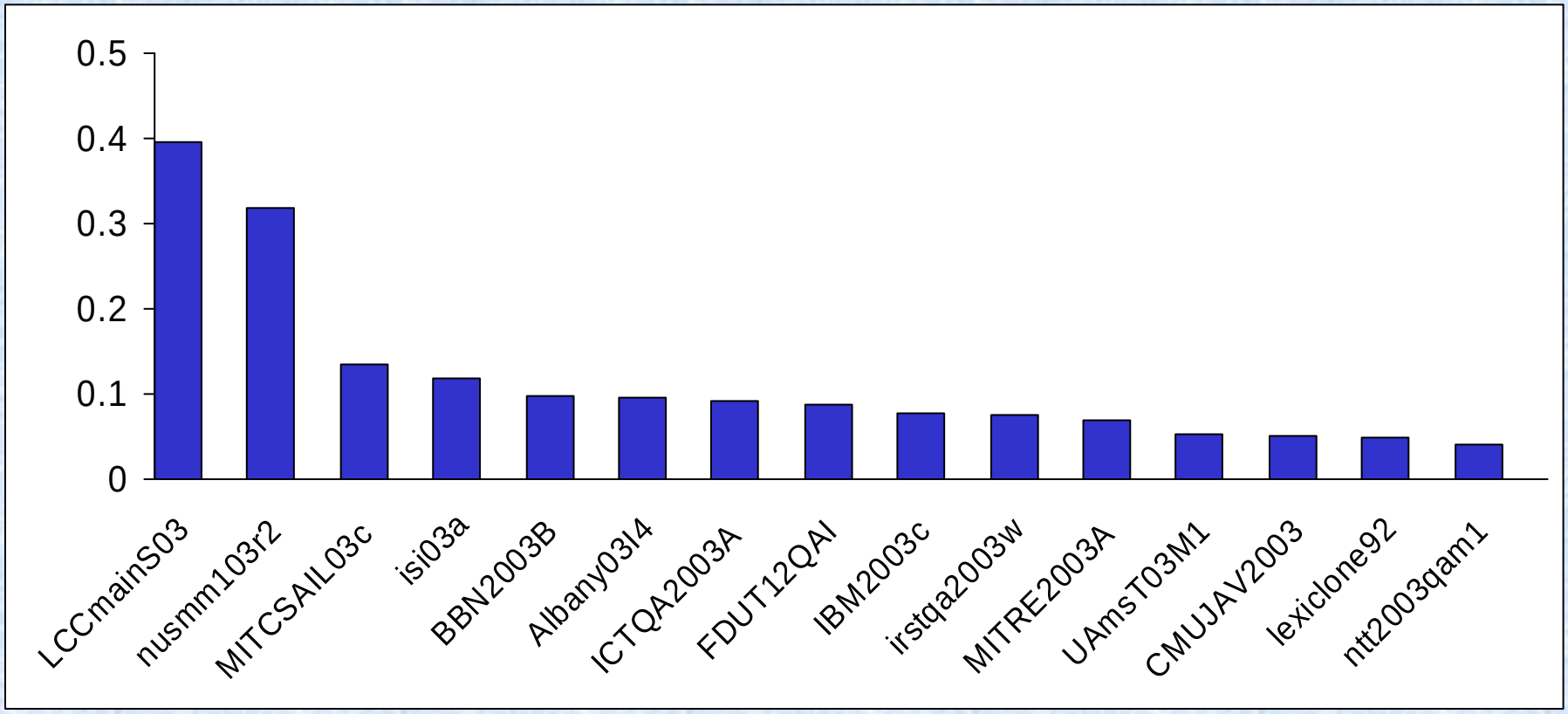
List Component of Main Task

- Systems return set of exact answers
  - questions seek multiple instances of a type
    - *List the names of chewing gums.*
    - *What Chinese provinces have a McDonald's restaurant?*
  - target number of instances not given
    - systems to return complete set of answers
    - multiple answers per doc and multiple docs with answers
- List scoring
  - single assessor per question
    - created list of known, correct answers from results
  - combine instance recall & precision using F

$$F = (2 \times P \times R)/(P+R)$$

  - average F for list questions is $\frac{1}{4}$ final score

Main Task List Results



F score of best run for top 15 groups for list component

Main Task Definition Component

- Definition questions
  - ask for definition of item or person
    - *Who is Vlad the Impaler?*
    - *What is Freddie Mac?*
    - *What is Ph in biology?*
  - very frequent question type in logs
- Assumed context
  - native English speaker
  - "average" adult reader of U.S. newspapers
  - found a term he/she wants more info on

Main Task Definition Component

- Have same "concept-matching" problem as in other NLP evals
  - want to reward systems for retrieving all of the important concepts required & penalize systems for retrieving irrelevant or redundant concepts
  - but concepts represented in English in many ways
  - no one-to-one correspondence between system items & concepts
- Different questions have very different numbers of required concepts

Definition Question Evaluation

- Have assessor create list of concepts that definition should contain
  - indicate essential "nuggets"
  - okay nuggets
- Mark nuggets in system responses
  - mark a nugget at most once
  - individual item may have multiple, one, or no nuggets

What is a golden parachute?

Assessor nuggets

1. Agreement between companies and top executives
2. Provides remuneration to executives who lose jobs
3. Remuneration is usually very generous
4. Encourages executives not to resist takeover beneficial to shareholders
5. Incentive for executives to join companies
6. Arrangement for which IRS can impose excise tax

Judged system response

2,3	a.	The arrangement, which includes lucrative stock options, a hefty salary, and a "golden parachute" if Gifford is fired
1	b.	Oh, Eaton has a new golden parachute clause in his contract
	c.	But some, including many of BofA's top executives joined the 216 and cashed in their "golden parachute" severance packages
6	d.	But if he quits or is dismissed during the 2 years after the merger, he will be paid $24.4 million, with Daimler-Chrysler paying the "golden parachute" tax for him and the taxes on the compensation paid to cover the tax.
4	e.	After the takeover, as jobs disappeared and BofA's stock tumbled, many saw him as a bumbler who sold out his bank, walking away with a golden parachute that gives him $5 million a year for the rest of his life.
	f.	The big payment that Eyler received in January was intended as a "golden parachute"

Quantitative Evaluation

- With this methodology, concept recall computable, but not concept precision
  - no satisfactory way to list all nuggets retrieved
  - assessors cannot enumerate all nuggets in text
    - granularity issue
    - unnatural task
- Rough approximation to nugget precision: length
  - count (non-white-space) characters in all items
  - intuition is that users prefer shorter of 2 definitions with same concepts

Scoring Function

numVitalMatches = # of vital nuggets retrieved

numVital = # of vital nuggets in list

numTotalMatches = # of essential & okay nuggets retrieved

C = character allowance per match (used 100)

$$\text{Recall} = \frac{\text{numVitalMatches}}{\text{numVital}}$$

Approximated Precision =

set okayLength = C×numTotalMatches

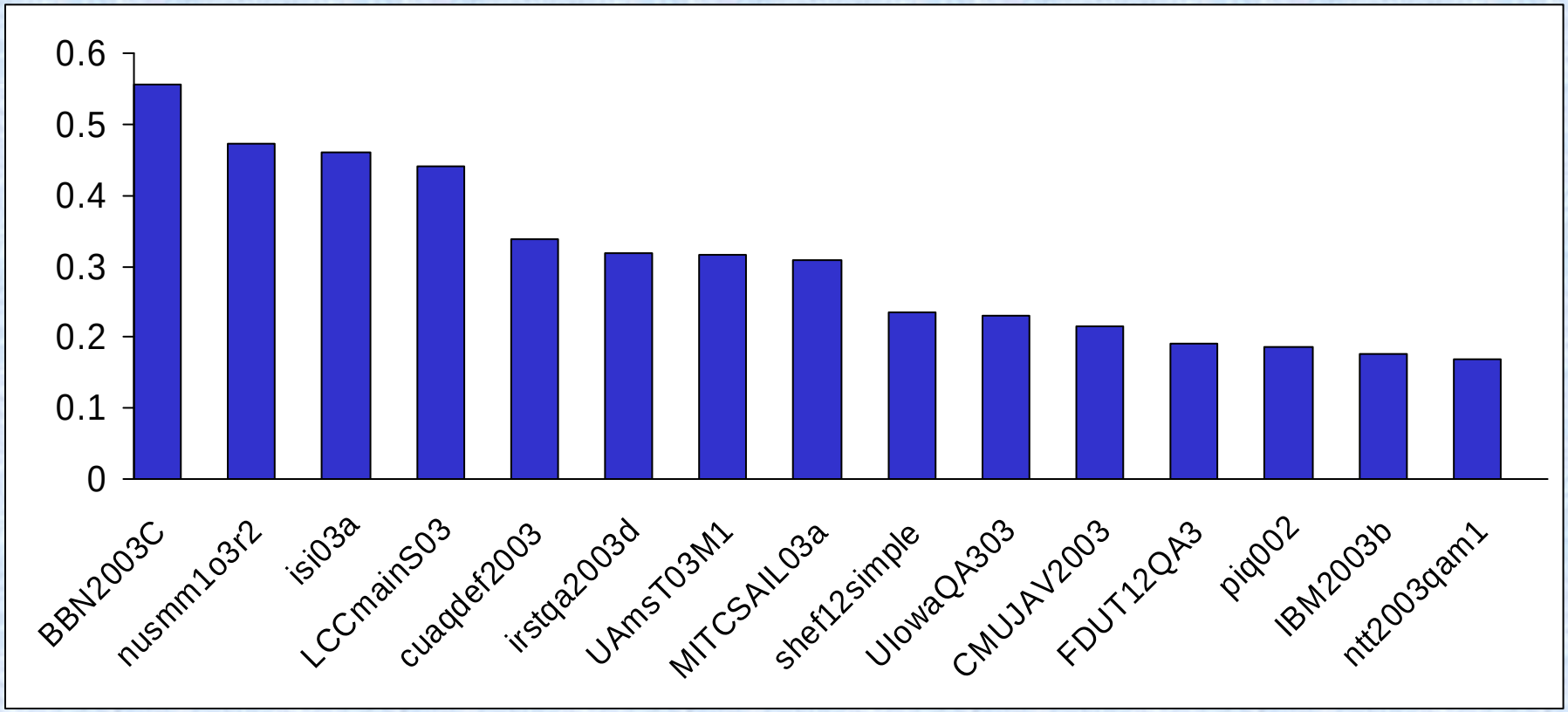
if (length < okayLength) then approxPrecision = 1

else approxPrecision = 1 –((length-okayLength)/length)

$$F = \frac{(\beta^2+1)RP}{(\beta^2 P+R)}$$

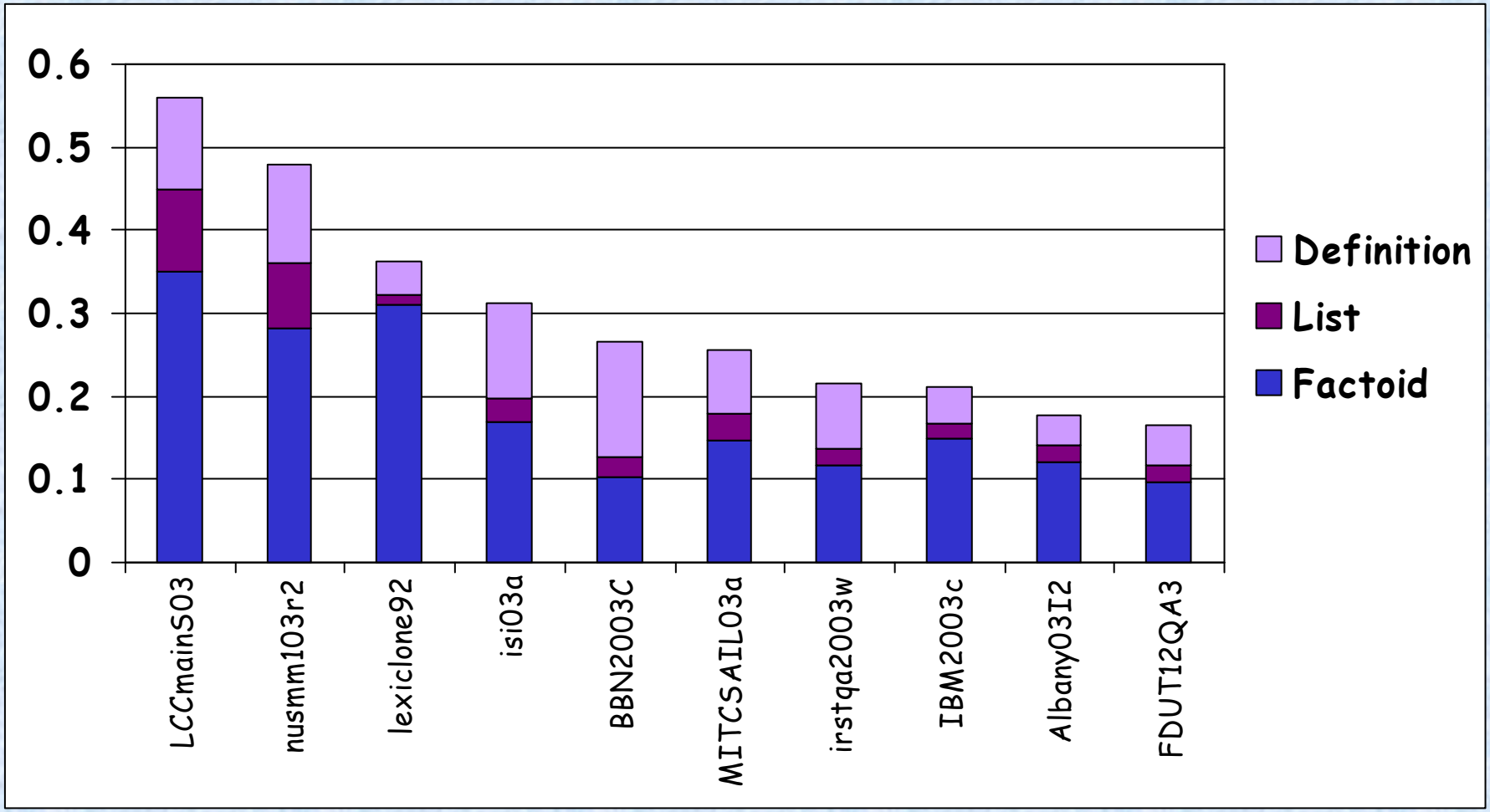
(used β=5)

Main Task Definition Results



F(β=5) score of best run for top 15 groups for definition component

QA Main Task Results

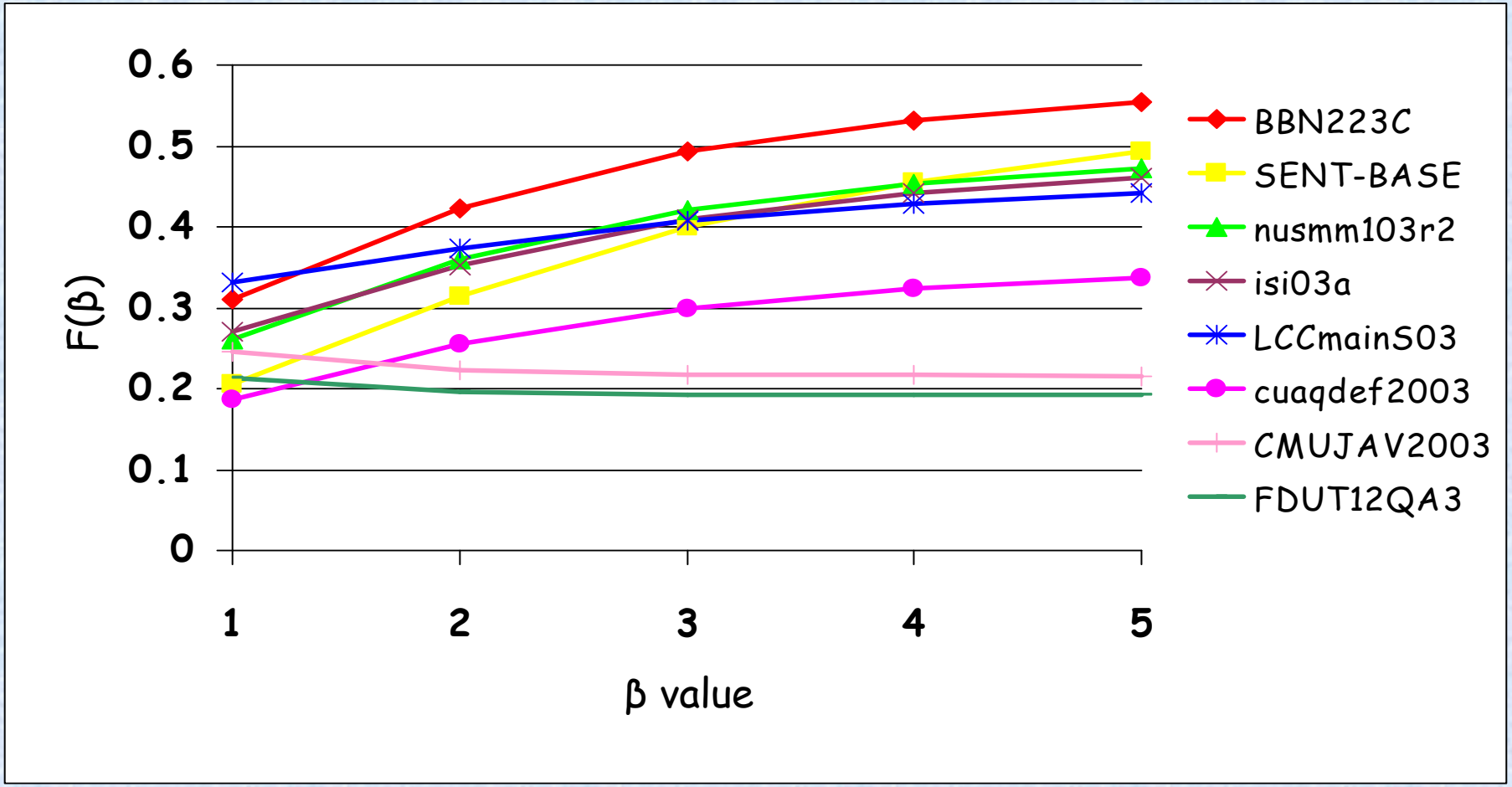


Final combined scores for best main task run per group for top 10 groups

Quality of the Definition Evaluation

- Two properties of an evaluation
  - fidelity
    - does evaluation measure appropriate thing?
    - depends on definition of (abstracted) evaluation task
  - reliability
    - can the results be trusted?
    - variety of sources of noise: mistakes, differences of opinion, particular sample of questions used
- Definition evaluation
  - fidelity
    - value of appropriate?
  - reliability
    - empirically examine effect of different sources of noise

Fidelity of Definition Evaluation



Effect of varying B on definition F scores; SENT-BASE is a baseline run that retrieves non-redundant sentences that mention target

Reliability of Definition Evaluation

- Mistakes by assessors
  - exist in all evaluations
  - can be directly measured since no pooling
- Differences of opinion
  - different assessors disagree as to correctness
  - inherent in NLP tasks
- Sample of questions
  - different systems do relatively differently on different questions
  - particular sample of questions can skew results
  - more questions lead to more stable results

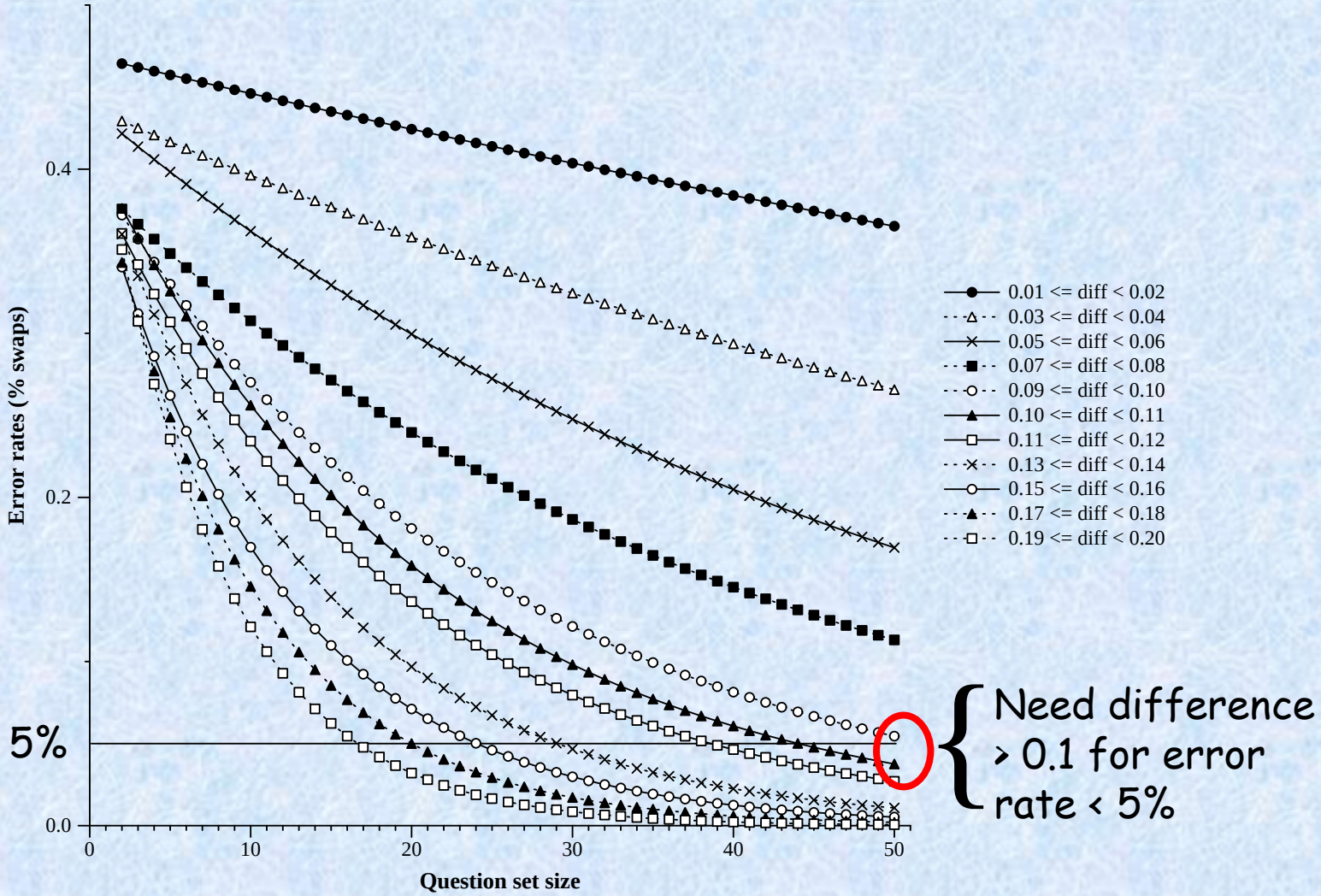
Mistakes by Assessors

- 14 pairs of identical definition components
  - across all pairs, 19 different definition questions judged differently
  - (roughly) uniform across assessors
  - number of questions affected ranges from 0 to 10
  - difference in F(β=5) scores ranges from 0.0 to 0.043, with a mean of 0.013
  - differences in F scores ≤ 0.043 for some different systems; clearly must be considered equivalent
- New task
  - consistency improved somewhat as assessors gained experience
  - better training re: granularity will help some
  - will never eliminate all errors

Differences of Opinion

- Each question independently judged by 2 assessors
  - assessors differed in what nuggets desired
  - assessors differed in whether nuggets vital
  - assessors did not differ as much in whether a nugget was present (modulo mistakes)
- Correlation among system rankings when questions judged by different assessors
  - compute Kendall τ correlation between rankings
  - τ=0.848, representing 113/1485 pairwise swaps
  - 8 swaps among systems whose F(β=5) scores as judged by original assessors differed by > 0.1
  - largest F(β=5) difference with swap was 0.123

Sample of Questions in Test Set



Definition Evaluation

- Noise within definition evaluation comparatively large
  - need to consider F scores within ± 0.1 of one another equivalent
  - coarse evaluation:
    - large equivalence classes of runs
  - one fix is to increase number of questions
    - larger sample of questions
    - individual mistakes have less effect
    - evaluation more costly