The RTE-3 Extended Task

Hoa Dang
Ellen Voorhees

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

History

- U.S. DTO AQUAINT program
  - focus on question answering for complex questions
  - long-standing interest in having systems justify their answers
  - RTE-3 provided opportunity to gather data from the community at large to discover what makes a good justification
- Extended task organization:
  - Chris Manning, Dan Moldovan, Ellen Voorhees
  - with thanks to the organizers of main RTE-3

Extended Task

- For each entailment pair in RTE-3 main task test set
  - provide 3-way entailment decision
    - Is Entailed (same as main task YES decision)
    - Contradicts (text and hypothesis cannot both be true)
    - Neither (the hypothesis neither is entailed by nor contradicts the text)
  - optionally, provide a justification for each decision where a justification is defined simply as a set of ASCII strings

3-way Entailment Decisions

- Motivation: drive systems to make more precise informational decisions
  - RTE main task conflates "can't tell" with "contradicts" when intuitively this is a big distinction
- Pragmatic definition for contradiction similar to that used in main task for entailment
  - based on "ordinary understanding"
  - assessors specifically told not to try to invalidate contradiction by constructing some bizarre interpretation where it might happen to be true

Examples

Text: The Communist Party USA was a small Maoist political party which was founded in 1965 by members of the Communist Party around Michael Laski who took the side of China in the Sino-Soviet split.

Hypothesis: *Michael Laski was an opponent of China.*

Decision: Contradicts

Text: Ms. Minton left Australia in 1961 to pursue her studies in London.

Hypothesis: *Ms. Hinton was born in Australia*

Decision: Neither entailed nor contradicts

3-Way Answer Key

- Constructed by NIST assessors
  - all 800 test pairs judged, with each pair judged by two different NIST assessors …
  - … but set of 'YES' decisions exactly the same as main task (regardless of NIST assessors' judgments)
  - disagreements between assessors for contradicts/ neither adjudicated by Ellen Voorhees
- Final statistics
  - 410 Entails
  - 72 Contradicts
  - 318 Neither

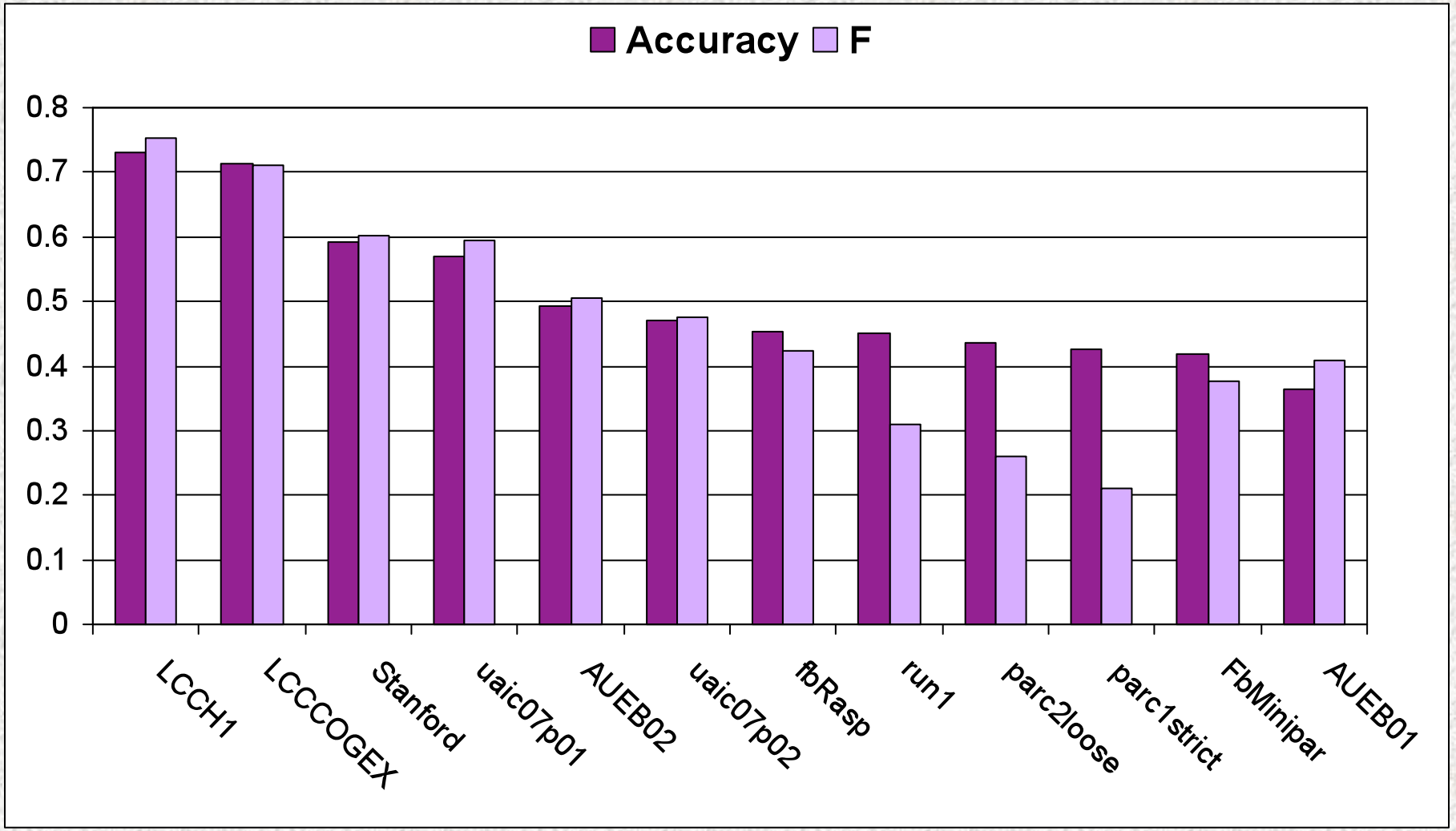
Evaluation

- Accuracy
  - percentage of system responses that match answer key
- $F = (1+\beta)\,(P*R) / (\beta*P+R)$
  - weighted harmonic mean of recall and precision
  - $\beta = 1/3$ (emphasize precision over recall)
  - computed only over 'Is Entailed' and 'Contradicts' entailment pairs
  - this combination selected to allow 'Neither' to be a reasonable 'Don't know' response, but highly skewed pair distribution since 410=YES, 72=NO

Submissions

Organization	Country	Run tags	Run type
Athens University of Economics and Business	Greece	AUEB01 AUEB02	3-way decisions only
Boeing	USA	run1	justifications included
LCC (COGEX)	USA	LCCCOGEX	justifications included
LCC (Groundhog)	USA	LCCH1	justifications included (manual run)
PARC	USA	parc1strict parc2loose	3-way decisions only
Stanford University	USA	Stanford	justifications included
University of Iasi	Romania	uaic07p01 uaic07p02	justifications included
University of Karlsruhe	Germany	fbRasp FbMinipar	3-way decisions only

Effectiveness of 3-way Decisions



True contradictions are rare and difficult for systems to recognize

Contingency table of responses over all entailment pairs and all runs

Systems say → Answer ↓ Key	Entails	Neither	Contradicts
Entails	2449	2172	299
Neither	929	2345	542
Contradicts	348	415	101

Note diagonal entry for 'Contradicts' substantially smallest entry in both row and column.

Contradictions are Hard (Take 2)

- No system had the correct response for 22 pairs
  - `Contradicts' correct for 20 pairs; `Entails' for other 2

9 (IE) Answer key: Contradicts

Text: Minton's first major part in England was as Maggie Dempster in the premiere of Nicholas Maw's One Man Show. Shortly thereafter, she became a regular member of the company of the Royal Opera House Covent Garden.

Hypothesis: *Maggie Dempster was a member of the company of the Royal Opera House Covent Garden.*

303 (IR) Answer key: Entails

Text: A settlement agreement between the federal government and the State of Florida, and approved by Judge William Hoeveler, imposed a plan to reduce damaging phosphorus levels in the Loxahatchee National Wildlife Refuge and Everglades National Park by December 31, 2006.

Hypothesis: *The US Government protected Everglades from further environmental damage.*

Diverse Responses from Systems

- All 12 systems agreed on only 2/800 pairs
  - all responded Entails when Entails is correct response

456 (QA)

Text: The Hubble is the only large visible-light and ultra-violet space telescope we have in operation.

Hypothesis: *Hubble is a Space telescope.*

501 (QA)

Text: Victor Emmanuel III (1869-1947) was king of Italy from 1900 to 1946. His cooperation with Mussolini helped bring an end to the Italian Monarchy.

Hypothesis: *Victor Emmanuel III was king of Italy from 1900 to 1946*.

Annotator Agreement

- Arbitrarily assign one NIST assessor as first assessor for a pair, the other as second
- Contingency tables show raw counts before adjudication
- Conflated agreement is percent of matches when both Neither and Contradict match NO

NIST → ↓ RTE	Entail	Neither	Contradict
YES	378	27	5
NO	48	242	100

RTE vs. one NIST meta-assessor; (conflated) agreement = .90

NIST → ↓ RTE	Entail	Neither	Contradict
YES	383	23	4
NO	46	267	77

RTE vs. other NIST meta-assessor; (conflated) agreement = .91

Example Disagreements

- Official key says Yes, both NIST Assessors said `Contradicts' (3-way key retains Yes)
  - Granularity of place names

Text: The 52nd Golden Globe Awards, honoring the best in film and television for 1994, were held on January 21, 1995 at the Beverly Hilton Hotel in Beverly Hills, California.

Hypothesis: *Golden Globes for 1994 were awarded in Los Angeles*.

  - Tense issues
    (NIST instructions to assessors differed from official)

Text: The head of the Israel Defense Forces, Lt. Gen. Dan Halutz resigned on Tuesday January 16, 2007 after an internal review criticized his military's leadership during the war in Lebanon last summer.

Hypothesis: *Dan Halutz is the head of the Israel Defense Forces.*

Annotator Agreement

- Agreement between two NIST assessors
- Table shows raw counts before adjudication
- Agreement is percentage of matching judgments

	Entail	Neither	Contradict
Entail	381	43	2
Neither	39	217	13
Contradict	9	30	66

NIST assessor vs. NIST assessor; agreement = .83

➢Level of agreement is consistent with other RTE challenges, but three-way (rather than binary) match means systems' accuracy scores can change by significant amount just by changing assessors

Justifications

- Motivation: explore what constitutes a good explanation of a system response
- Guidelines explicitly noted that target was “ordinary” end user
  - specifically not system builder (not debugging output!)
  - user not necessarily linguist or logician
- Specification of allowable justifications deliberately vague: “a collection of ASCII strings”
  - don’t want to arbitrarily preclude good ideas in pilot
  - but was text-based rather than graphical
  - no size limit imposed; submitted justifications brief

Example

Text: Muybridge had earlier developed an invention he called the Zoopraxiscope.

Hypothesis: *The Zoopraxiscope was invented by Muybridge*.

LCCCOGEX	LCCH1	Stanford
The text mentions 'Muybridge'. We can infer that Muybridge is inventor. From the fact that Muybridge is inventor, we can infer that Muybridge invented. We can conclude that the Zoopraxiscope was invented by Muybridge.	There is a relationship between Zoopraxiscope and Muybridge in both the text and hypothesis. The term "invention" is morphologically similar to "invented".	1. The Hypothesis could be precisely matched with content in the Text, with allowance for polarity and embedded context. 2. Hypothesis words match well with words in the Text. 3. text adjunct "called" of "invention" dropped on aligned hyp word "invented"

Example (cont'd)

Text: Muybridge had earlier developed an invention he called the Zoopraxiscope.

Hypothesis: *The Zoopraxiscope was invented by Muybridge*.

run1

Yes! I have general knowledge that:
IF Y is developed by X THEN Y is manufactured by X
Here: X = Muybridge, Y = the invention
Thus, here:
We are told in t: the invention is developed by Muybridge
Thus it follows that: the invention is manufactured by Muybridge
In addition, I know "manufacture" and "invent" mean roughly the same thing

Hence: the Zoopraxiscope was invented by Muybridge.

uaic07p01

The words in the hypothesis are all found, with approximately all the same syntactic dependencies, also in the text. Therefore, I concluded that given hypothesis:

The Zoopraxiscope was invented by Muybridge

is entailed by the given text

Muybridge had earlier developed an invention he called the Zoopraxiscope.

Judging Justifications

- Given: an entailment pair, the answer key response, a system's response & justification
  - assign a score on a scale of 1(low) – 5 (high) for how *understandable* the justification is
  - if the understandability is at least 3, assign a score on a scale of 1 (low) – 5 (high) for how *compelling* the argument contained in the justification is; otherwise, no compellingness score is assigned
    - compellingness involves both correctness and pertinence
    - a system reaching the wrong conclusion could get partial credit for a compellingness score, but not a '5'

Judging Justifications

- Assessors could skip pairs for which they disagreed with the answer key response
- Assessors judged all justifications for a given entailment pair before moving on to new pair
- Presentation order of different systems' justifications randomized for a given entailment pair; all assessors who judged that pair saw same order

Judging Justifications

- NIST (Ellen) selected 100 pairs as subset of justifications to be judged
  - various factors in selection: systems generally reached correct response for pair; balance across four application areas (IE, IR, QA, Sum); both ‘long’ and ‘short’ examples; the justification for at least one system was `interesting’ in some way
- All six NIST assessors judged all 100 pairs
- Recall NIST received 6 runs that included justifications from 5 groups; one run was manual

Scores Assigned to Example

Text: Muybridge had earlier developed an invention he called the Zoopraxiscope.

Hypothesis: *The Zoopraxiscope was invented by Muybridge*.

LCCCOGEX

The text mentions 'Muybridge'. We can infer that Muybridge is inventor. From the fact that Muybridge is inventor, we can infer that Muybridge invented. We can conclude that the Zoopraxiscope was invented by Muybridge.

scores:

[4 3] [3 3] [5 4]
[5 1] [5 3] [3 2]

LCCH1

There is a relationship between Zoopraxiscope and Muybridge in both the text and hypothesis. The term "invention" is morphologically similar to "invented".

scores:

[4 4] [4 4] [5 4]
[4 1] [5 4] [3 2]

[3 3] [4 4] [4 4]
[2 -] [1 -] [2 -]

Stanford

1. The Hypothesis could be precisely matched with content in the Text, with allowance for polarity and embedded context.

2. Hypothesis words match well with words in the Text.

3. text adjunct "called" of "invention" dropped on aligned hyp word "invented"

Example (cont'd)

run1

Yes! I have general knowledge that:
IF Y is developed by X THEN Y is manufactured by X
Here: X = Muybridge, Y = the invention

Thus, here:
We are told in t: the invention is developed by Muybridge
Thus it follows that: the invention is manufactured by Muybridge

In addition, I know "manufacture" and "invent" mean roughly the same thing

Hence: the Zoopraxiscope was invented by Muybridge.

scores:

[2 -] [4 1] [3 3] [3 1] [2 -] [1 -]

uaic07p01

The words in the hypothesis are all found, with approximately all the same syntactic dependencies, also in the text. Therefore, I concluded that given hypothesis:

The Zoopraxiscope was invented by Muybridge

is entailed by the given text

Muybridge had earlier developed an invention he called the Zoopraxiscope.

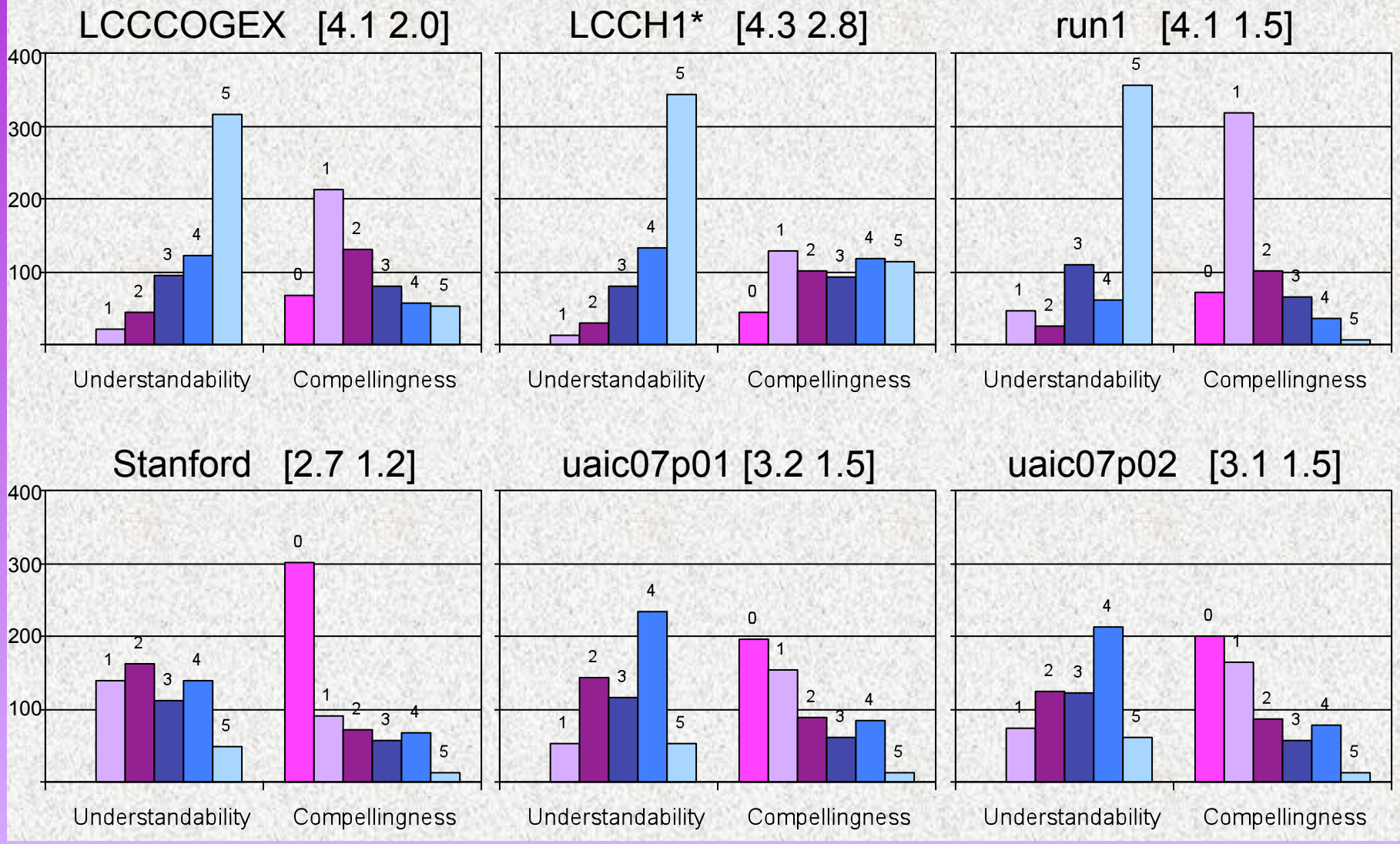
scores:

[3 3] [4 3] [4 3] [5 1] [4 3] [2 -]

Scoring

- Lots of numbers!
  - each run: 100 pairs X 6 assessors X 2 5-point scores
  - participants received complete set of 1200 numbers for their run
  - computed mean understandability score and mean compellingness score
    - average computed per assessor over 100 pairs, per entailment pair over 6 assessors, and grand overall mean
    - mean actually a dubious choice since 5-point scale should probably be treated as category rather than interval variable
    - if pair was unjudged, it was skipped in computing means; if understandability was <3, compellingness treated as ‘0’

Distribution of Justification Scores

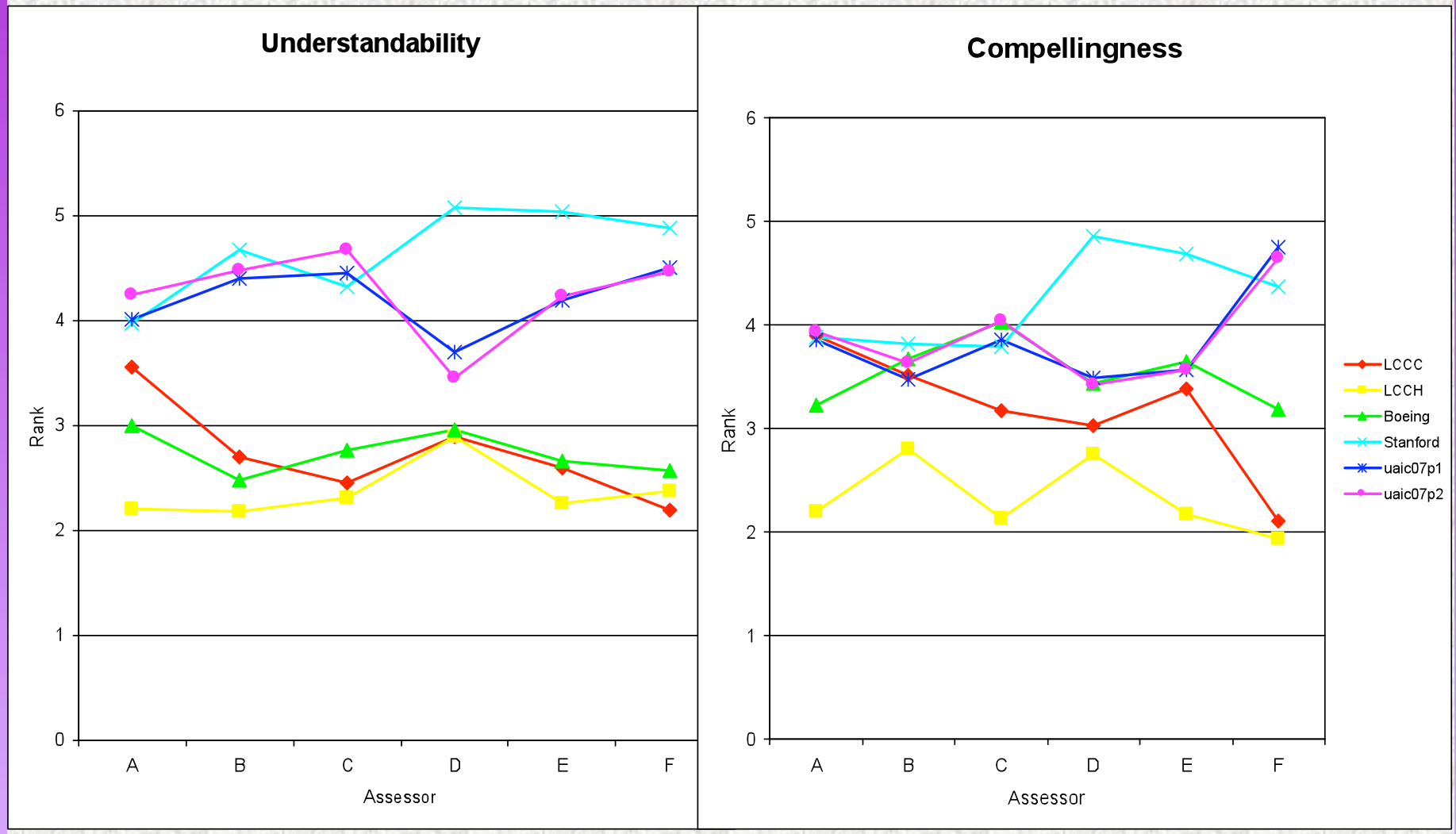


NIST

Assessor Agreement

- Differences for both Understandability and Compellingness
  - even with separate score, compellingness score still affected by ease of comprehension
  - the six NIST assessors have different mathematical backgrounds, and it showed
  - differences do affect system rankings

Average System Rank



Common Themes

- Conciseness highly prized
  - "chatty" explanations grew old fast
  - want explanations that focus on specifics
    - "There is a relation", "there is a match" are unsatisfying
  - but mathematical notation also largely discouraged
- Don't discuss system internals
  - (uncalibrated) scores from various components do not install confidence
  - jargon (polarity, adjunct, …) unpopular

Extended Pilot Results

- 3-way decisions
  - good task; need more test pairs for evaluation stability
  - recognizing actual contradiction currently difficult
  - contradiction relatively rare in 2007 test set, so constructing future balanced test set may be an issue
- Justifications
  - need better understanding of overall purpose of task
    - having system explain itself at this level is probably counterproductive
    - need to appeal to true task for effective use of assessors
  - systems reaching correct entailment decision by faulty reasoning uncomfortably often