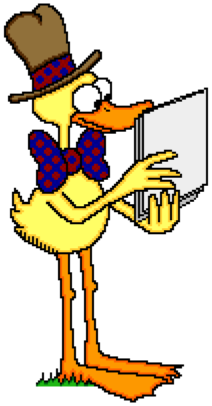


An Introduction to DUC-2003

Intrinsic Evaluation of Generic News Text Summarization Systems

Paul Over
Retrieval Group
Information Access Division

James Yen
Statistical Modeling and Analysis Group
Statistical Engineering Division

National Institute of Standards and Technology

Sponsored by DARPA and ARDA

Document Understanding Conferences (DUC)…

- Summarization has always been a TIDES component
- An evaluation roadmap created in 2000 after spring TIDES PI meeting
- Specifies a series of annual cycles
- Year 1 (DUC-2001 at SIGIR in September 2001)
  - Intrinsic evaluation of generic summaries,
    - of newswire/paper stories
    - for single and multiple documents;
    - with fixed target lengths of 50, 100, 200, and 400 words
  - 60 sets of 10 documents used
    - 30 for training
    - 30 for test

… Document Understanding Conferences (DUC)

- Year 2 – short cycle – (DUC-2002 at ACL ’02 in July 2002)
  - Intrinsic evaluation of generic summaries,
    - of newswire/paper stories
    - for single and multiple documents
  - Abstracts of single documents and document sets
    - fixed lengths of 10, 50, 100, and 200 words
    - manual evaluation using SEE software at NIST
  - Extracts of document sets
    - fixed target lengths of 200 and 400 words
    - automatic evaluation at NIST and by participants
  - 60 sets of ~10 documents each
    - All for test
    - No new training data
    - Two abstracts/extracts per document (set)

Goals of the talk

- Provide an overview of DUC 2003:
  - Data: documents, topics, viewpoints, manual summaries
  - Tasks:
    - 1: very short (~10-word) single document summaries
    - 2-4: short (~100-word) multi-document summaries with focus

      2: TDT event topics

      3: viewpoints

      4: question/topic
  - Evaluation: procedures, measures
    - Experience with implementing the evaluation procedure
- Introduce the results (what happened):
  - Basics of system performance on the measures
  - Sanity checking the results and measures
  - Exploration of various questions:
    - Performance of systems relative to baselines and humans
    - Relative performance among systems – significant differences?

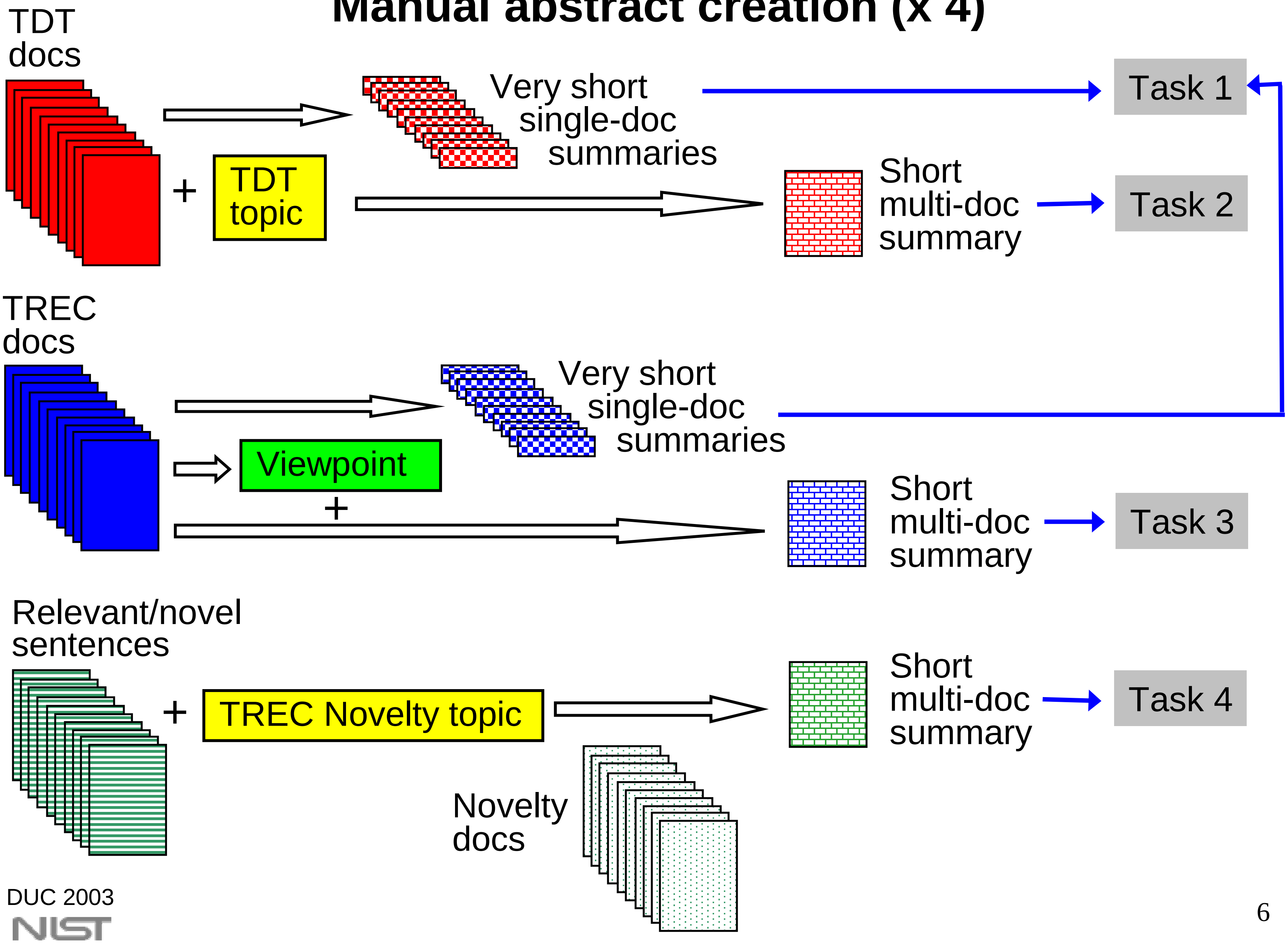
Data: Formation of test document sets

- 30 TDT clusters (298 documents; ~352 sentences/docset)
  - 30 event topics and documents chosen by NIST
    - 15 from TDT2
    - 15 from TDT3
  - NIST chose a subset of the documents the TDT annotator decided were "on topic"

- 30 TREC clusters (326 documents; ~335 sentences/docset)
  - Chosen by NIST assessors on topics of interest to them
  - No restrictions as to topic type

- 30 TREC Novelty clusters (~66 relevant sentences/docset)
  - 30 Novelty topics picked by NIST (based on assessor agreement)
  - All (~25) Novelty track documents/cluster included
  - Relevant/novel sentences identified by Novelty assessors

Manual abstract creation (x 4)



TDT docs

+ TDT topic

Very short single-doc summaries

Task 1

Short multi-doc summary

Task 2

TREC docs

Viewpoint

+

Very short single-doc summaries

Short multi-doc summary

Task 3

Relevant/novel sentences

+ TREC Novelty topic

Novelty docs

Short multi-doc summary

Task 4

Baseline summaries etc.

- NIST (Nega Alemayehu) created baseline summaries
  - Baselines 2-5: automatic
  - based roughly on algorithms suggested by Daniel Marcu
  - no truncation of sentences, so some baseline summaries went over the limit (+ <=15 words) and some were shorter than required)
- Original author’s headline 1 (task 1)
  - Use the document’s own “headline” element
- Baseline 2 (tasks 2, 3)
  - Take the 1st 100 words in the most recent document.
- Baseline 3 (tasks 2, 3)
  - Take the 1st sentence in the 1st, 2nd, 3rd,… document in chronological sequence until you have 100 words.
- Baseline 4 (task 4)
  - Take the 1st 100 words from the 1st n relevant sentences in the 1st document in the set. (Documents ordered by relevance ranking given with the topic.)
- Baseline 5 (task 4)
  - Take the 1st relevant sentence from the 1st, 2nd, 3rd,… document until you have 100 words. (Documents ordered by relevance ranking given with the topic.)

Submitted summaries by system and task

SYSID	Code	T1	T2	T3	T4	Group
AMDS_HW.v1	6	-	30	-	-	Heriot-Watt University
uam.duc2003.v6	7	624	-	-	-	University of Madrid
gistkey.duc03	8	624	-	-	-	Federal U. of Sao Carlos
bbn.umd.hedge	9	624	-	-	-	BBN / U. of Maryland
CL.Research.duc03	10	622	30	30	30	CL Research
cslab.duc03	11	-	30	30	-	NTT
fudan.duc2003	12	-	30	-	-	Fudan University
isiwebcl.duc2003.vcombined	13	624	30	30	30	ISI/USC
aquaintandmultigenanddems	14	-	30	-	30	Columbia University
ku.duc2003	15	624	30	30	-	Korea University
ccsnsa.duc03.v3	16	-	30	30	29	NSA+
UofLeth-DUC2003	17	624	30	30	30	University of Lethbridge
kul.2003	18	624	30	30	-	University of Leuven
SumUMFAR	19	-	30	-	30	University of Montreal
crl_nyu.duc03	20	-	30	30	30	New York University
uottawa	21	624	30	30	-	University of Ottowa
lcc.duc03	22	624	30	30	30	LCC
UofM-MEAD	23	-	30	30	30	University of Michigan
UDQ	24	564	-	-	-	University of Girona
CLaC.DUCTape.Summarizer	25	624	-	-	-	Concordia University
saarland.2003	26	624	30	-	-	Univ. of the Saarland

Evaluation basics

- Content coverage and linguistic quality:
  - Intrinsic evaluation by humans using special rewritten version of SEE (thanks to Lei Ding and Chin-Yew Lin at ISI)
  - Compare:
    - a model summary - authored by a human
    - a peer summary - system-created, baseline, or additional manual
  - Produce judgments of:
    - Peer quality (12 questions)
    - Coverage of each model unit by the peer (recall)
    - Relevance of peer-only material

- Usefulness (task 1) and Responsiveness (task 4):
  - Simulated extrinsic evaluations
  - Comparison together of all peer summaries for a given doc(set)
  - Assignment of each summary to one of 5 bins

Models

- Source:
  - Authored by a human
  - For 2003, the assessor is always the model's author
- Formatting:
  - Divided into model units (MUs)
    - (MUs == EDUs - thanks to Radu Soricut at ISI)
  - Lightly edited by authors to integrate uninterpretable fragments
    - George Bush's selection of Dan Quale
    - as his running mate surprised many
    - many political observers thought him a lightweight with baggage
    - to carry
  - Flowed together with HTML tags for SEE

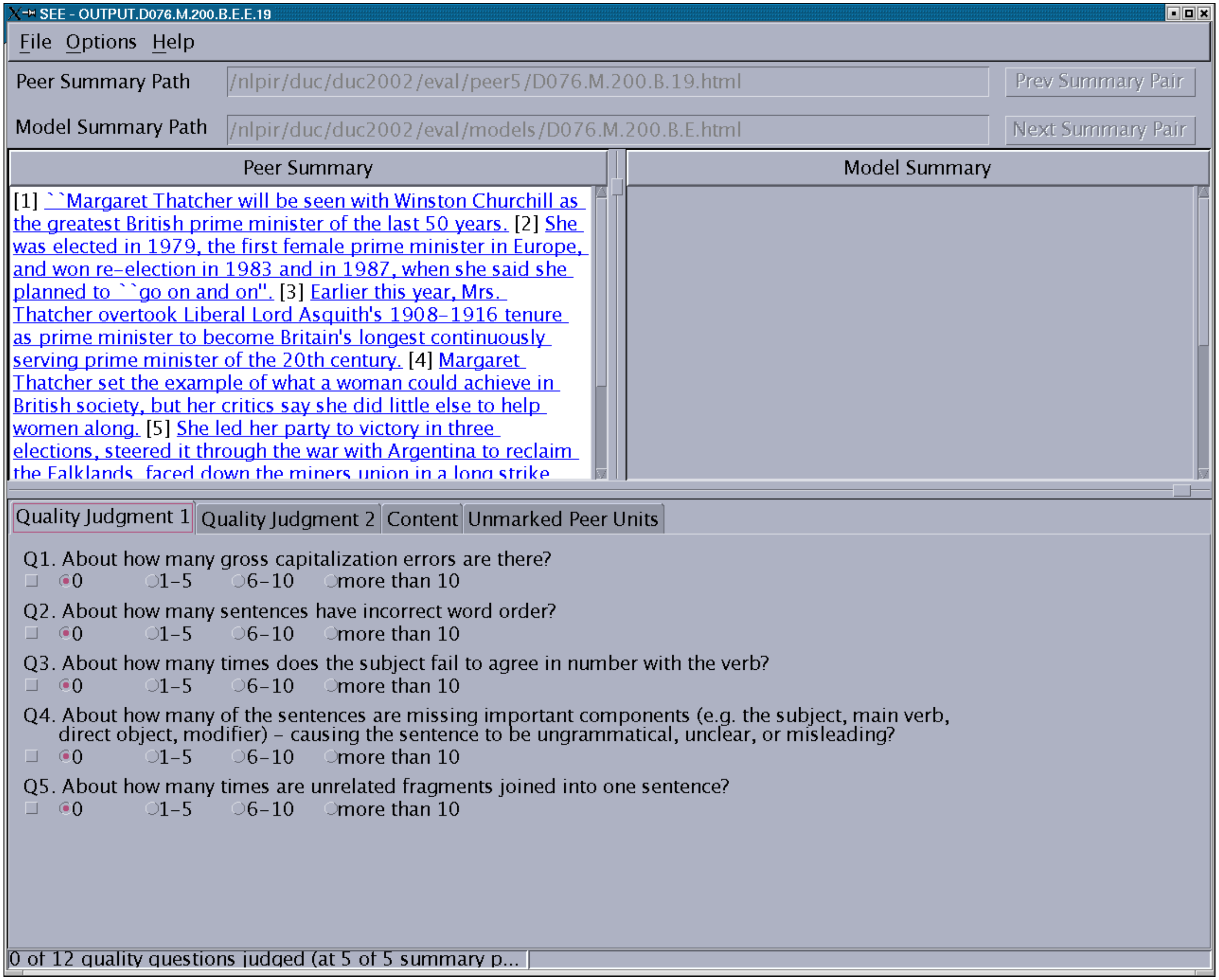
Peers

- Formatting:
  - – Divided into peer units (PUs) –
    - simple automatically determined sentences
    - tuned slightly to documents and submissions
      - – Abbreviations list
      - – List of proper nouns
    - Flowed together with HTML tags for SEE

- 4 Sources:

1. Author’s headline: 1
2. Automatically generated by baseline algorithms: 2 – 5
3. Automatically generated by research systems: 6 – 26
4. Authored by a human other than the assessor: A – J

SEE: overall peer quality



Peer Summary Path /nlpir/duc/duc2002/eval/peer5/D076.M.200.B.19.html

Model Summary Path /nlpir/duc/duc2002/eval/models/D076.M.200.B.E.html

Peer Summary

[1] ``Margaret Thatcher will be seen with Winston Churchill as the greatest British prime minister of the last 50 years. [2] She was elected in 1979, the first female prime minister in Europe, and won re-election in 1983 and in 1987, when she said she planned to ``go on and on''. [3] Earlier this year, Mrs. Thatcher overtook Liberal Lord Asquith's 1908-1916 tenure as prime minister to become Britain's longest continuously serving prime minister of the 20th century. [4] Margaret Thatcher set the example of what a woman could achieve in British society, but her critics say she did little else to help women along. [5] She led her party to victory in three elections, steered it through the war with Argentina to reclaim the Falklands, faced down the miners union in a long strike

Model Summary

Q1. About how many gross capitalization errors are there?
□ ●0 ○1-5 ○6-10 ○more than 10

Q2. About how many sentences have incorrect word order?
□ ●0 ○1-5 ○6-10 ○more than 10

Q3. About how many times does the subject fail to agree in number with the verb?
□ ●0 ○1-5 ○6-10 ○more than 10

Q4. About how many of the sentences are missing important components (e.g. the subject, main verb, direct object, modifier) - causing the sentence to be ungrammatical, unclear, or misleading?
□ ●0 ○1-5 ○6-10 ○more than 10

Q5. About how many times are unrelated fragments joined into one sentence?
□ ●0 ○1-5 ○6-10 ○more than 10

0 of 12 quality questions judged (at 5 of 5 summary p...

Overall peer quality

12 Questions developed with participants

Answer categories: 0 1-5 6-10 >10

1. About how many gross capitalization errors are there?
2. About how many sentences have incorrect word order?
3. About how many times does the subject fail to agree in number with the verb?
4. About how many of the sentences are missing important components (e.g. the subject, main verb, direct object, modifier) – causing the sentence to be ungrammatical, unclear, or misleading?
5. About many times are unrelated fragments joined into one sentence?

Overall peer quality

6. About how many times are articles (a, an, the) missing or used incorrectly?
7. About how many pronouns are there whose antecedents are incorrect, unclear, missing, or come only later?
8. For about how many nouns is it impossible to determine clearly who or what they refer to?
9. About how times should a noun or noun phrase have been replaced with a pronoun?
10. About how many dangling conjunctions are there ("and", "however"...)?
11. About many instances of unnecessarily repeated information are there?
12. About how many sentences strike you as being in the wrong place because they indicate a strange time sequence, suggest a wrong cause-effect relationship, or just don't fit in topically with neighboring sentences?

Overall peer quality

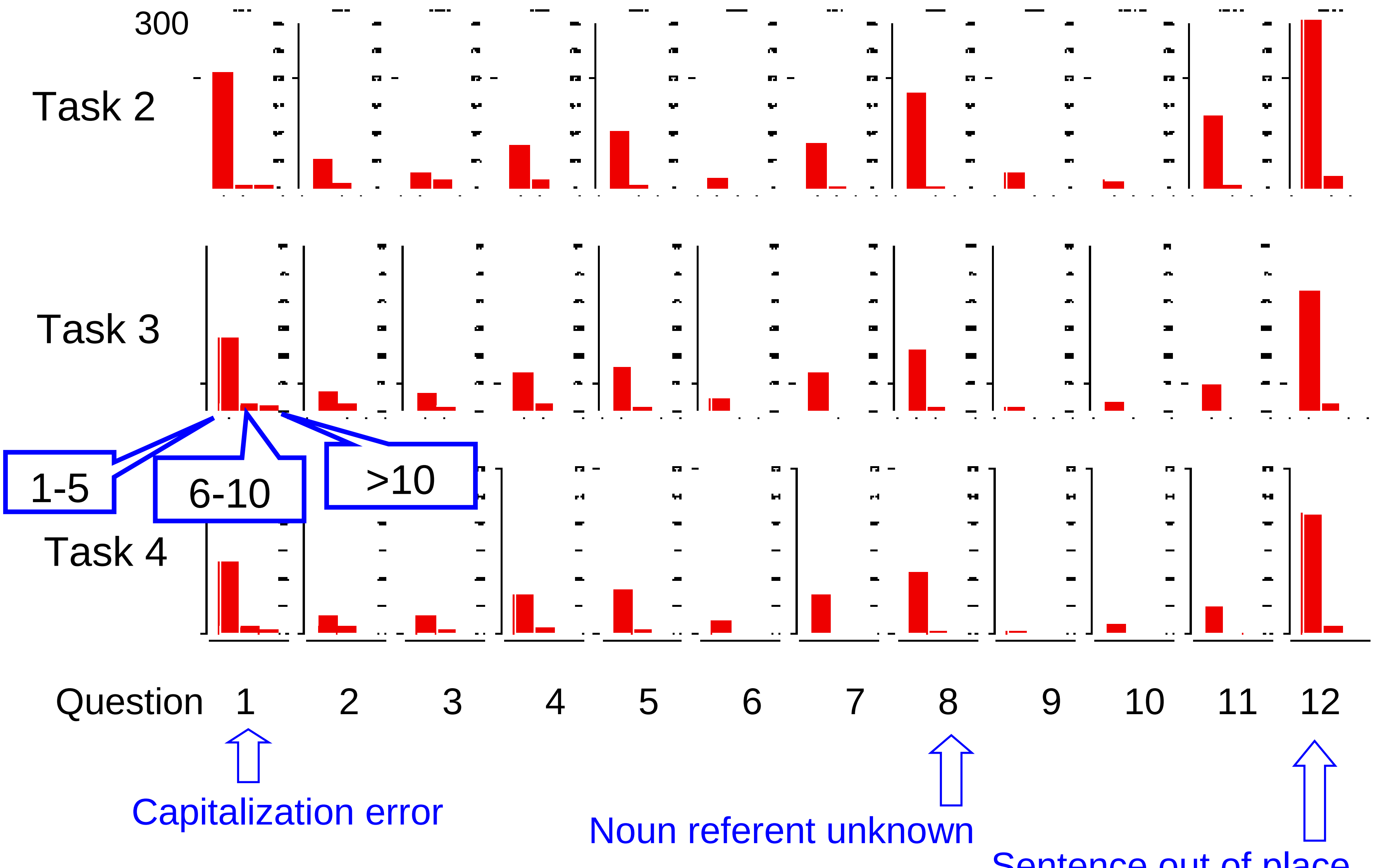
Systems > Baselines >= Manual

Mean number of quality questions indicating one or more errors

	n	Mean	~95% CI	Max
Task 2				
Systems	450	2.379	2.189 - 2.569	10
Baselines	60	0.900	0.786 - 1.014	3
Manuals	90	0.622	0.442 - 0.882	5
Task 3				
Systems	330	2.315	2.108 - 2.522	9
Baselines	60	1.048	0.935 - 1.161	3
Manuals	90	0.356	0.207 - 0.505	4
Task 4				
Systems	269	1.963	1.772 - 2.154	9
Baselines	60	0.742	0.616 - 0.868	2
Manuals	89	0.386	0.221 - 0.551	3

Overall peer quality

Uneven distribution of non-zero scores by question

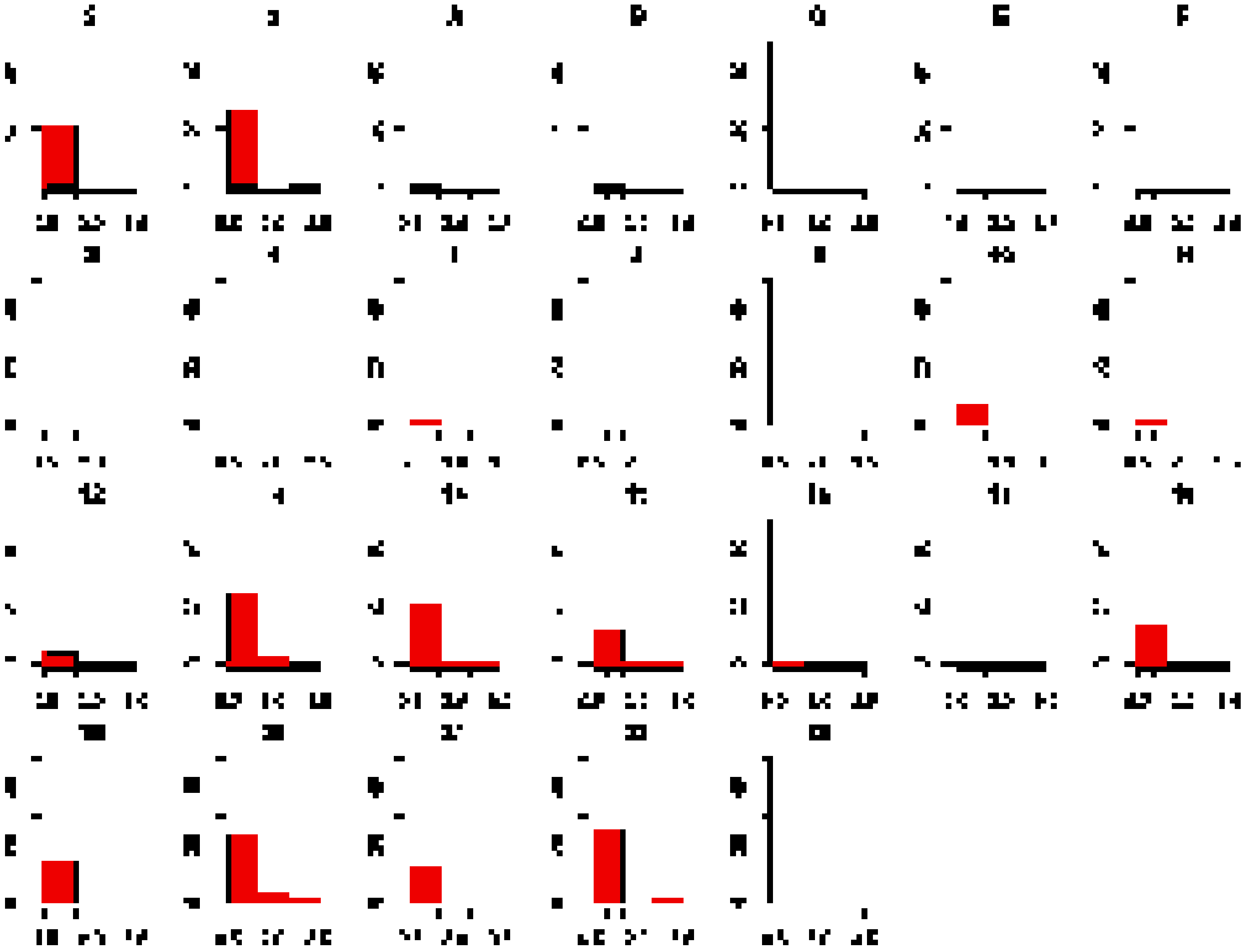


Capitalization error

Noun referent unknown

Sentence out of place

Overall peer quality
Q1: Capitalization



Overall peer quality

Q1: Capitalization

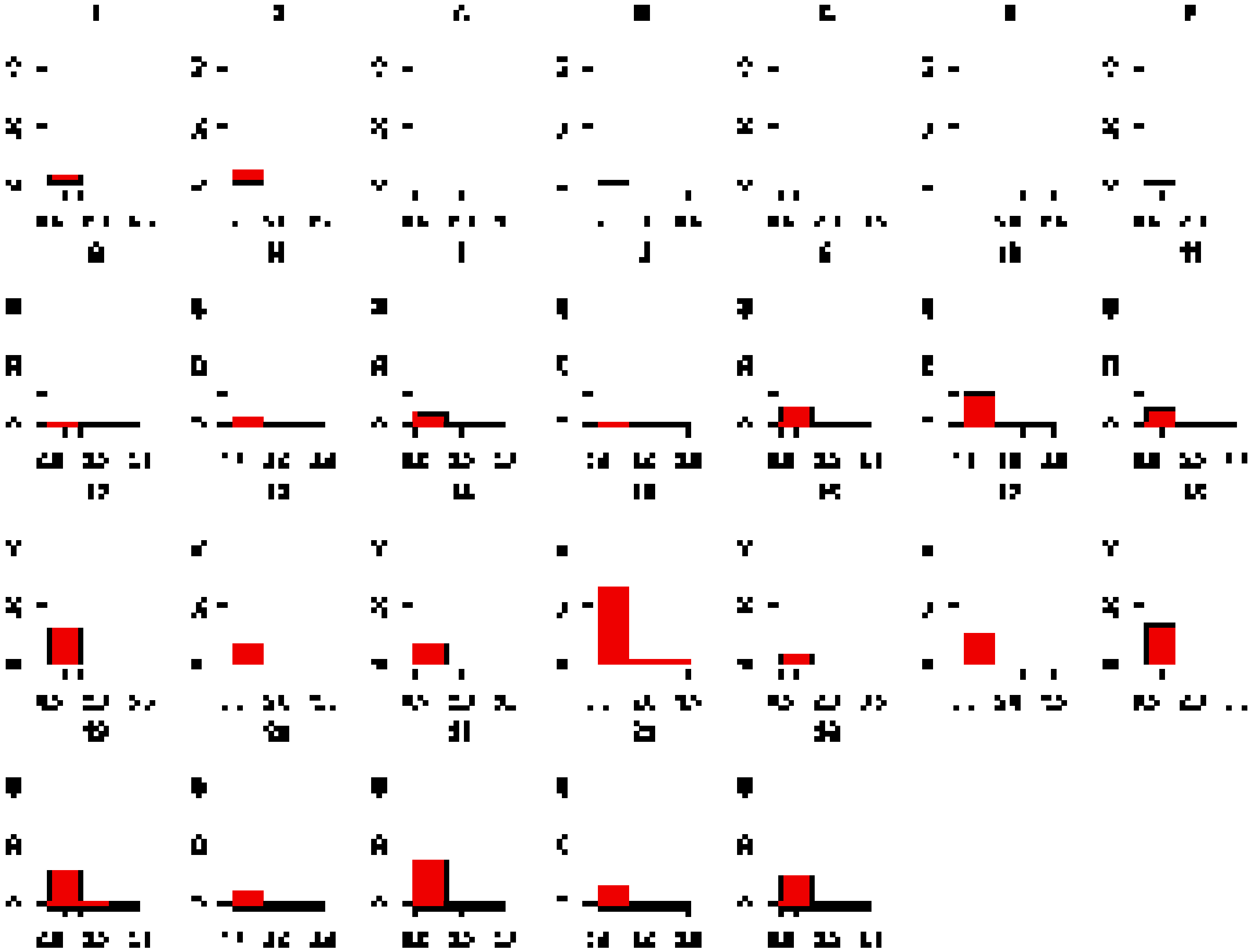
PARIS, February 20 (Xinhua) -- Declaring that "Currency is politics," French Prime Minister Alain Juppe today reiterated France's determination to realize the single European currency.

LONDON, March 28 (Xinhua) -- British officials will fight suggestions that UK be forced to enter a new European exchange rate mechanism (ERM) after the proposed European single currency comes into force, it was reported here today.

LONDON, April 4 (Xinhua) -- British Board of Trade president Ian Lang Wednesday warned that a single European currency could prove harmful to British business if adopted without full and careful consideration of possible consequences.

Overall peer quality

Q8: Noun referent unknown



DUC 2003

Overall peer quality

Q8: Noun referent unknown

The president indicate that he is willing to strip some of the anti-environmental he wrote that impact his state riders.

That $18 billion on the International Monetary Fund spending bes a waste of money convince conservatives.

Dick Armey R-Texas did not predict that the GOP presence in Congress would be even stronger next year when the deal might be reached.

Republicans attach the president to deem to be anti-environment provisions.

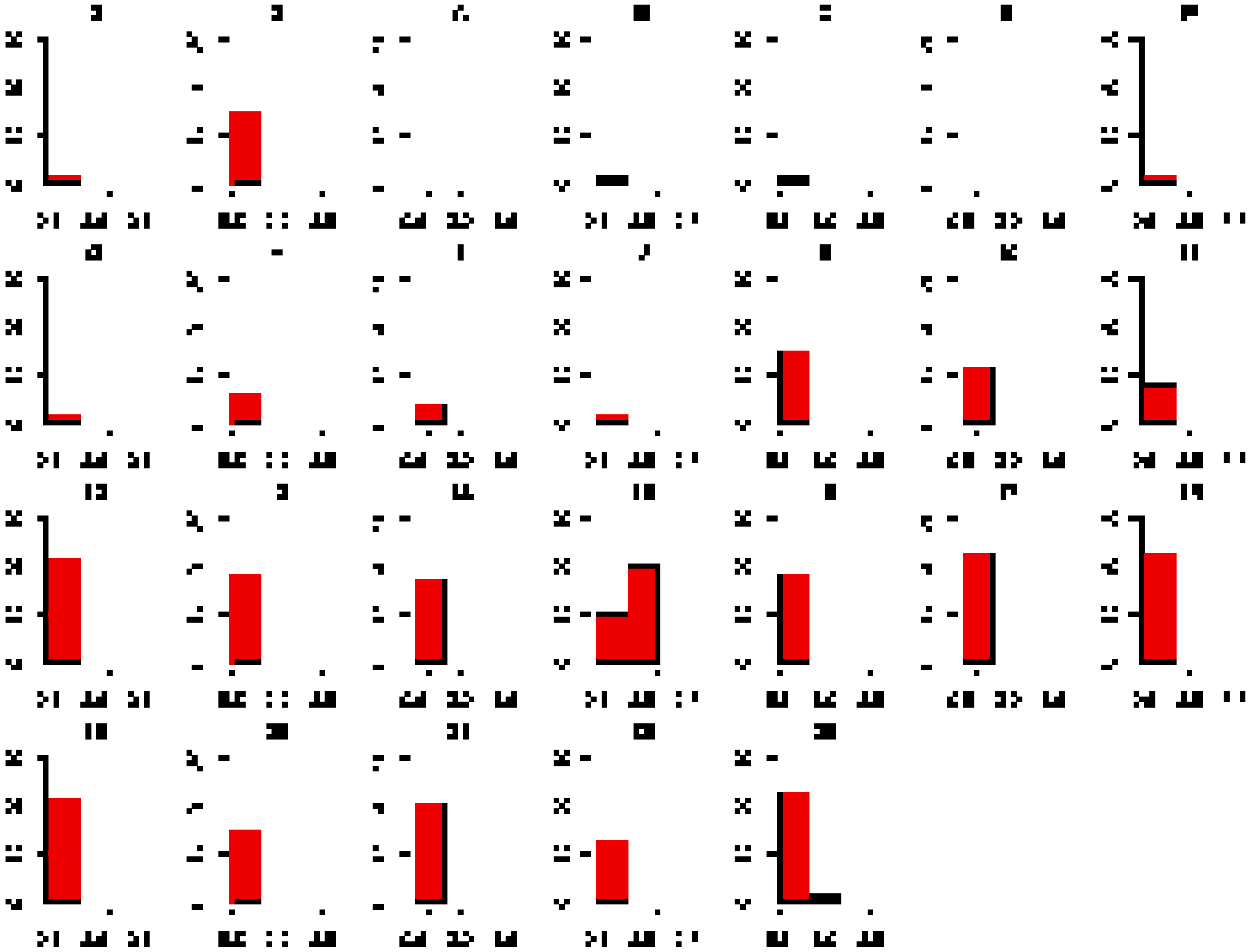
You know We 're that they are about a domestic thinking concerned.

Everybody understand the IMF can have American tax dollars.

The White House ever have that until mid-September.

Overall peer quality

Q12: Misplaced sentences



Overall peer quality
Q12: Misplaced sentence(s)

All of these satellites came through Tuesday's meteor shower unscathed.

Showers of Leonid meteors may produce hundreds or thousands of blazing meteors each hour.

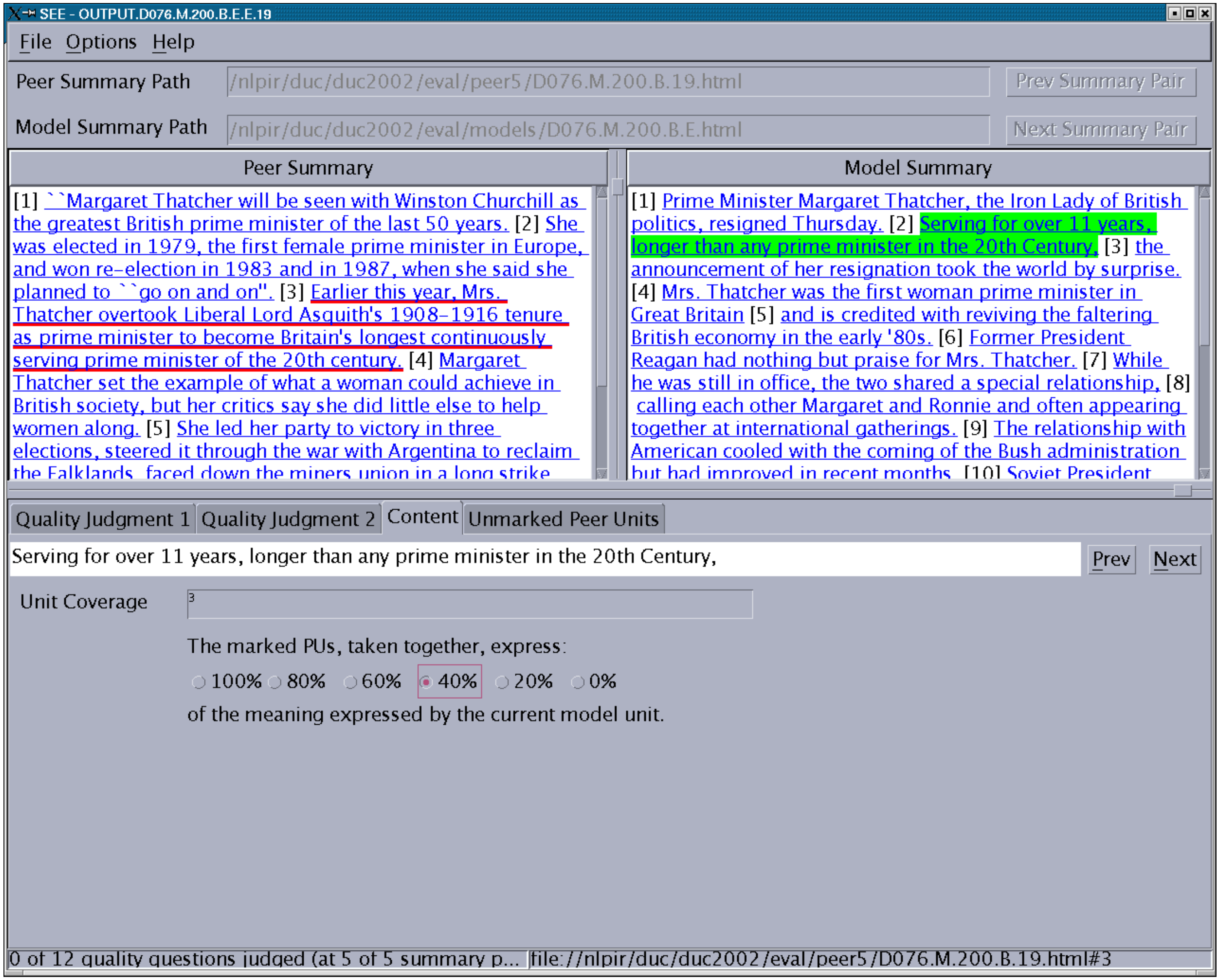
Some satellites in low-earth orbits can actually hide from meteoroid storms, Ozkul said.

The scientists who track Temple-Tuttle do not even call it a shower, they call it a meteor storm.

Satellite experts said that some damage might take days to detect, but that satellites generally seemed to have escaped disabling harm.

This storm of meteors, called Leonid meteors because they come from the direction of constellation Leo, will be the first to hit the Earth since 1966 when the world's space programs were in their infancy, and its effects on satellite systems are uncertain.

SEE: per-unit content



Peer Summary Path /nlpir/duc/duc2002/eval/peer5/D076.M.200.B.19.html

Model Summary Path /nlpir/duc/duc2002/eval/models/D076.M.200.B.E.html

Peer Summary

[1] ``Margaret Thatcher will be seen with Winston Churchill as the greatest British prime minister of the last 50 years. [2] She was elected in 1979, the first female prime minister in Europe, and won re-election in 1983 and in 1987, when she said she planned to ``go on and on''. [3] Earlier this year, Mrs. Thatcher overtook Liberal Lord Asquith's 1908-1916 tenure as prime minister to become Britain's longest continuously serving prime minister of the 20th century. [4] Margaret Thatcher set the example of what a woman could achieve in British society, but her critics say she did little else to help women along. [5] She led her party to victory in three elections, steered it through the war with Argentina to reclaim the Falklands, faced down the miners union in a long strike

Model Summary

[1] Prime Minister Margaret Thatcher, the Iron Lady of British politics, resigned Thursday. [2] Serving for over 11 years, longer than any prime minister in the 20th Century, [3] the announcement of her resignation took the world by surprise. [4] Mrs. Thatcher was the first woman prime minister in Great Britain [5] and is credited with reviving the faltering British economy in the early '80s. [6] Former President Reagan had nothing but praise for Mrs. Thatcher. [7] While he was still in office, the two shared a special relationship, [8] calling each other Margaret and Ronnie and often appearing together at international gatherings. [9] The relationship with American cooled with the coming of the Bush administration but had improved in recent months. [10] Soviet President

Serving for over 11 years, longer than any prime minister in the 20th Century,

Unit Coverage 3

The marked PUs, taken together, express:

○ 100% ○ 80% ○ 60% ● 40% ○ 20% ○ 0%

of the meaning expressed by the current model unit.

0 of 12 quality questions judged (at 5 of 5 summary p... file://nlpir/duc/duc2002/eval/peer5/D076.M.200.B.19.html#3

Per-unit content: evaluation details

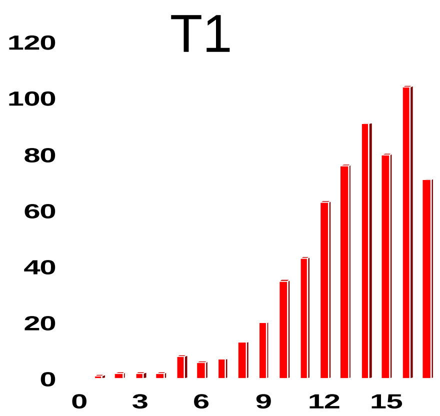
- “First, find all the peer units which tell you at least some of what the current model unit tells you, i.e., peer units which express at least some of the same facts as the current model unit. When you find such a PU, click on it to mark it.
- Requirement for common facts relaxed for very short summaries
  - Common references count
- “When you have marked all such PUs for the current MU, then think about the whole set of marked PUs and answer the question:”
- “The marked PUs, taken together, express about
  [0%   20%   40%   60%   80%   100%]
  of the meaning expressed by the current model unit”

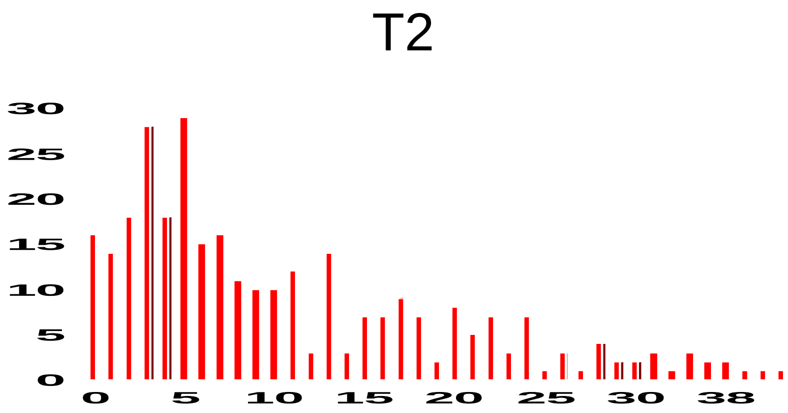
Per-unit content:
% MU-to-peer comparisons with no coverage

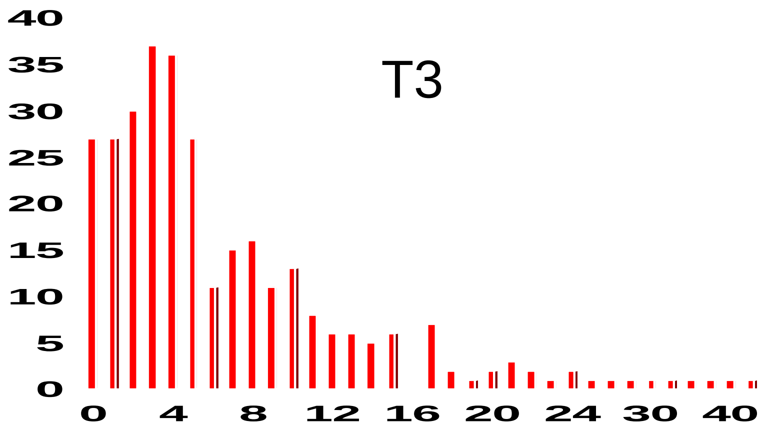
	All	Manual	Automatic
Task 1	20.9	5	24.3
Task 2	60.3	35.1	64.5
Task 3	68.9	48.6	73.6
Task 4	67.9	45.7	73.9
Task 4*	66	44.2	71.9

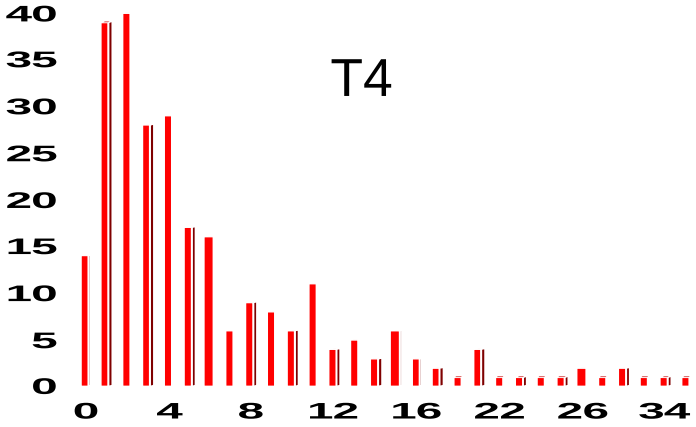
- DUC 2002:
  - All - 62%
  - Manual – 42%
- DUC 2001
  - All - 63%
- Appear to be due to real differences in content
- Do the peers agree on which MUs are not covered?

Per-unit content: Counts of MUs by number of PUs mapped to them









Per-unit content measures: – recall

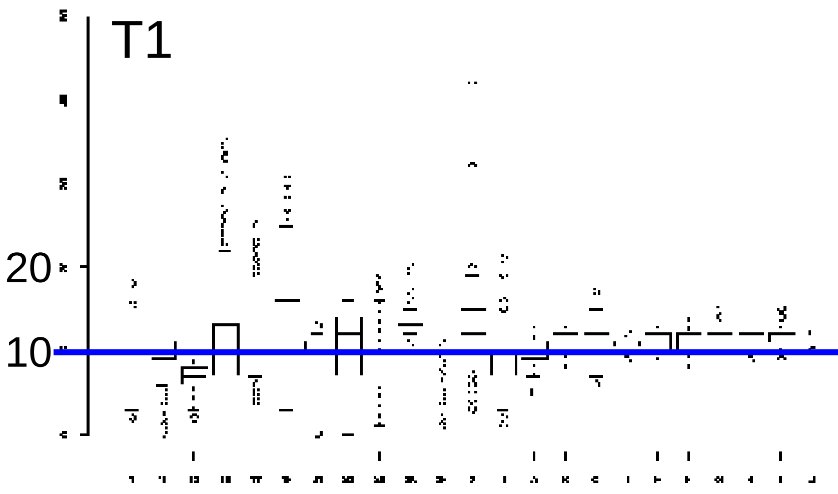
- What fraction of the model content is also expressed by peer?
- Mean coverage:
  - average of the per-MU completeness judgments [0, 20, 40, 60, 80,100]% for a peer summary
- Mean length-adjusted coverage (2002):
  - average of the per-MU length-adjusted coverage judgments for a peer
  - length-adjusted coverage = 2/3 * coverage + 1/3 * brevity where brevity =
    - 0 if actual summary length >= target length; else
    - (target size – actual size) / target size

Ø Sets two goal: complete coverage and smallest possible summary

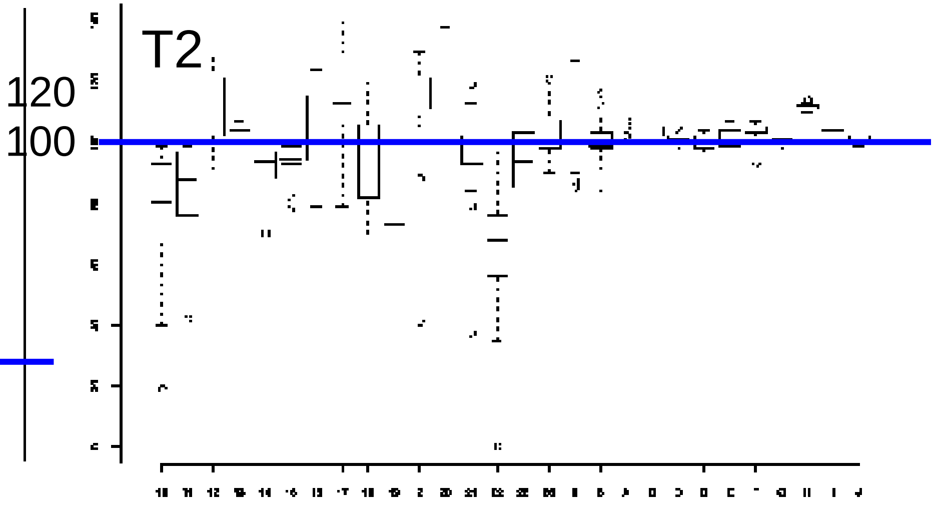
Ø Perfect score only possible when BOTH goals reached

Ø Truncate if target size exceeded

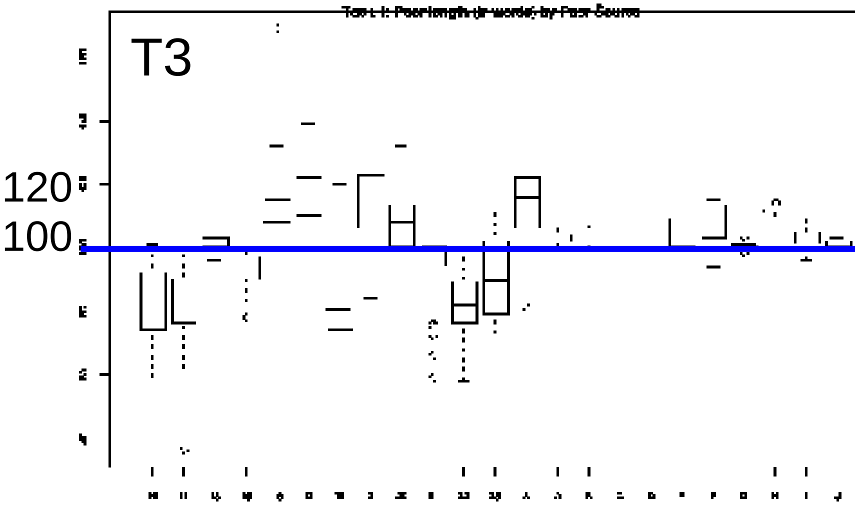
Summary lengths (in words)
by peer



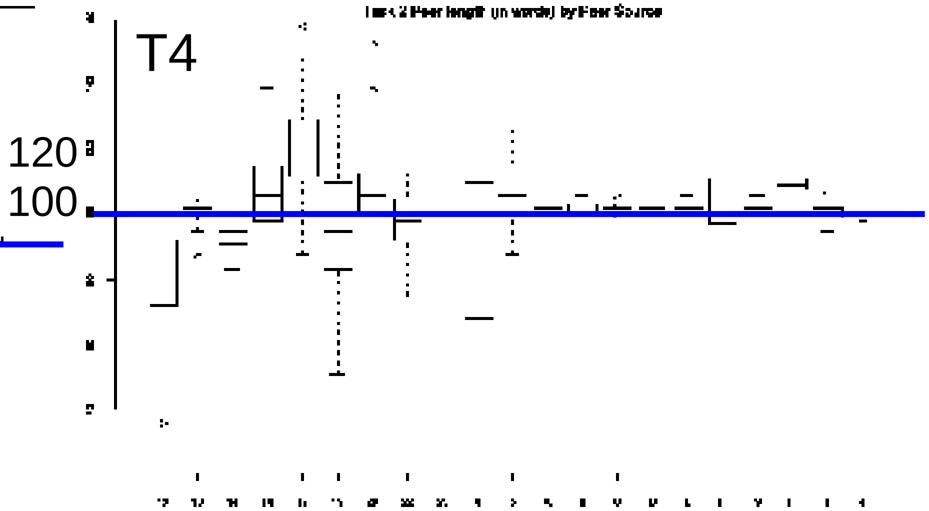
Task 1: Peer length (in words) by Peer Source



Task 2 Peer length (in words) by Peer Source



Task 3 Peer length (in words) by Peer Source



Task 4 Peer length (in words) by Peer Source

Per-unit content measures: – recall

- Task 1: Coverage
  - coverage
  - coverage with penalty iff over target length

    = coverage * target size / actual size

    - **Post hoc substitute for lack of truncation**

- Tasks 2-4: Length-adjusted coverage (LAC)
  - improved

    coverage = 0 à LAC = 0

  - Improved, with penalty iff over target length

    = LAC * target size / actual size

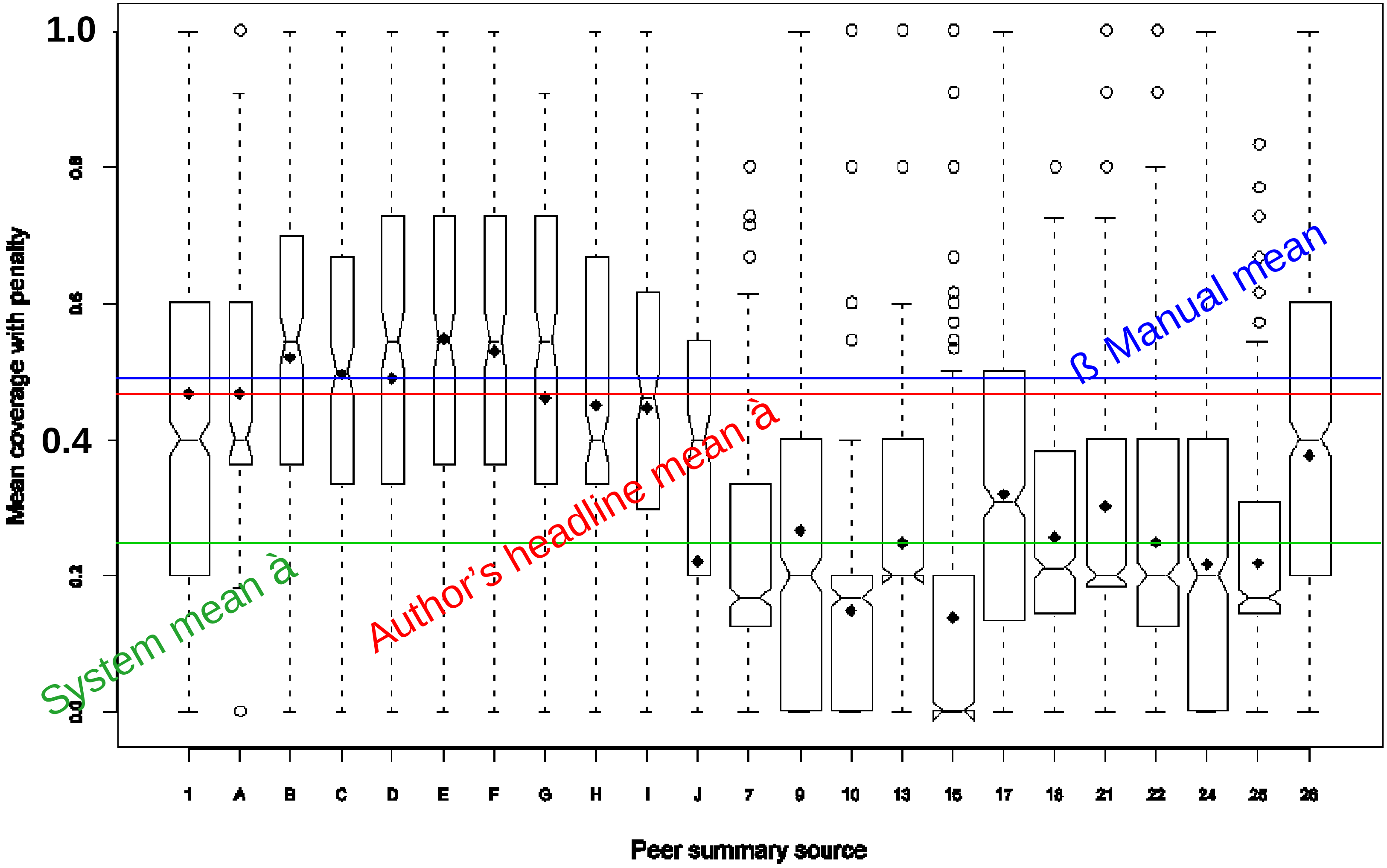
  - proportional

    = coverage * target size / actual size

Task 1: Very short summary of a single document

- System task:
  - Use the 30 TDT clusters and the 30 TREC clusters
    - 734 documents;
    - ~12 documents/cluster
  - Given:
    - Each document
  - Create a very short summary
    - (~10 words, no specific format other than linear) of it.
- Evaluation:
  - SEE
    - Coverage
    - Extra material
  - Usefulness

Task 1: Mean coverage with penalty by peer



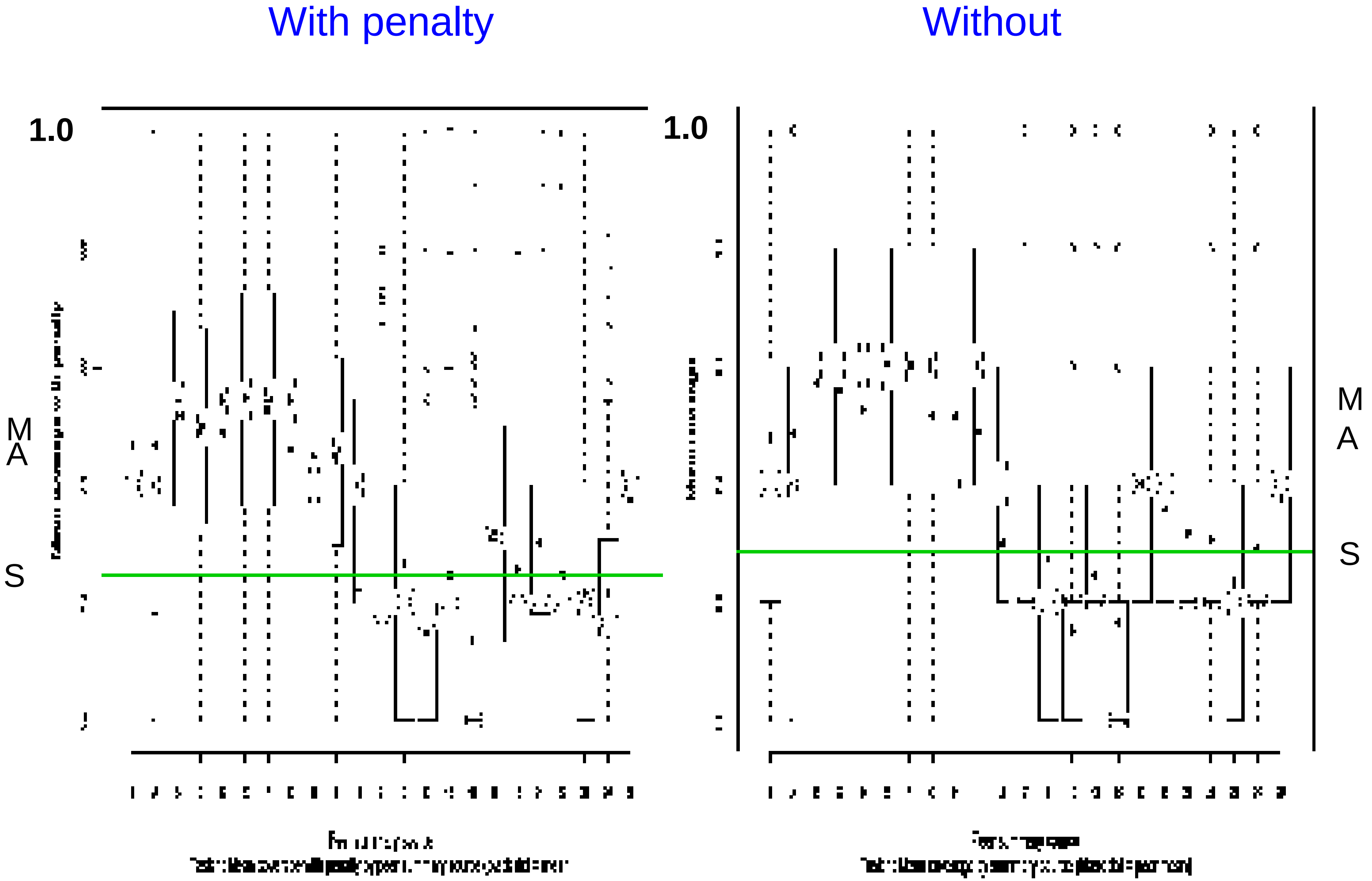
ß Manual mean

Author’s headline mean à

System mean à

Task 1: Mean coverage with penalty by peer summary source (black dot = mean)

Task 1: Mean coverage +/-penalty by peer



With penalty

Without

Task 1: ANOVA
(mean coverage with penalty)

```
Number of observations    9922
 The GLM Procedure

        R-Square     Coeff Var      Root MSE       Mean
        0.297547      67.80859      0.208265       0.307137

Source                      DF       Type I SS    Mean Square   F Value
docset                      59      42.1070990      0.7136796     16.45
peer                        22     138.6796453      6.3036202    145.33

                  Source                Pr > F
                  docset                <.0001
                  peer                  <.0001
```

Task 1: Multiple comparisons

(@ 0.05 confidence level)

SAS REGWQ Grouping		Mean	N	peer
	A	0.47981	624	1
	B	0.40160	624	17
C	B	0.37788	624	26
C		0.35801	624	18
	D	0.31763	624	21
	D	0.30609	624	22
	D	0.30000	624	7
	D	0.29199	624	25
E	D	0.27468	624	9
E		0.24744	624	13
E		0.23511	564	24
	F	0.16603	624	15
	F	0.15338	622	10

Coverage

			Mean	N	peer
	A		0.46712	624	1
	B		0.37686	624	26
	C		0.32009	624	17
	C		0.30272	624	21
	D		0.26770	624	9
E	D		0.25560	624	18
E	D	F	0.24923	624	22
E	D	F	0.24744	624	13
E		F	0.22206	624	7
		F	0.21866	624	25
		F	0.21750	564	24
	G		0.14949	622	10
	G		0.13825	624	15

Coverage with penalty

Means with the same letter are not significantly different.

Task 1: Usefulness

- Simulated extrinsic evaluation
- Assessor sees
  - each document
  - all summaries of that document
- Assessor asked to:
  - “Assume the document is one you should read.”
  - “Grade each summary according to how useful you think it would be in getting you to choose the document:

    0 (worst, of no use), 1, 2, 3, or 4 (best)”
- Double assessment

Task 1: Usefulness – Examples

[Document NYT20000415.0068 text]

4 U D107.P.10.C.H.H.A.NYT20000415.0068 :: **False convictions turn some conservatives against death penalty.**

1 U D107.P.10.C.H.H.7.NYT20000415.0068 :: **[death] their views seem incompatible; a number have raised; The columnist George Will wrote that skepticism.**

4 U D107.P.10.C.H.H.1.NYT20000415.0068 :: **LOOK WHO'S QUESTIONING THE DEATH PENALTY**

3 U D107.P.10.C.H.H.J.NYT20000415.0068 :: **Conservatives, death penalty, morality, DNA, justice, Will, Pat Robertson, Republican**

0 U D107.P.10.C.H.H.9.NYT20000415.0068 :: **ranks are admittedly small**

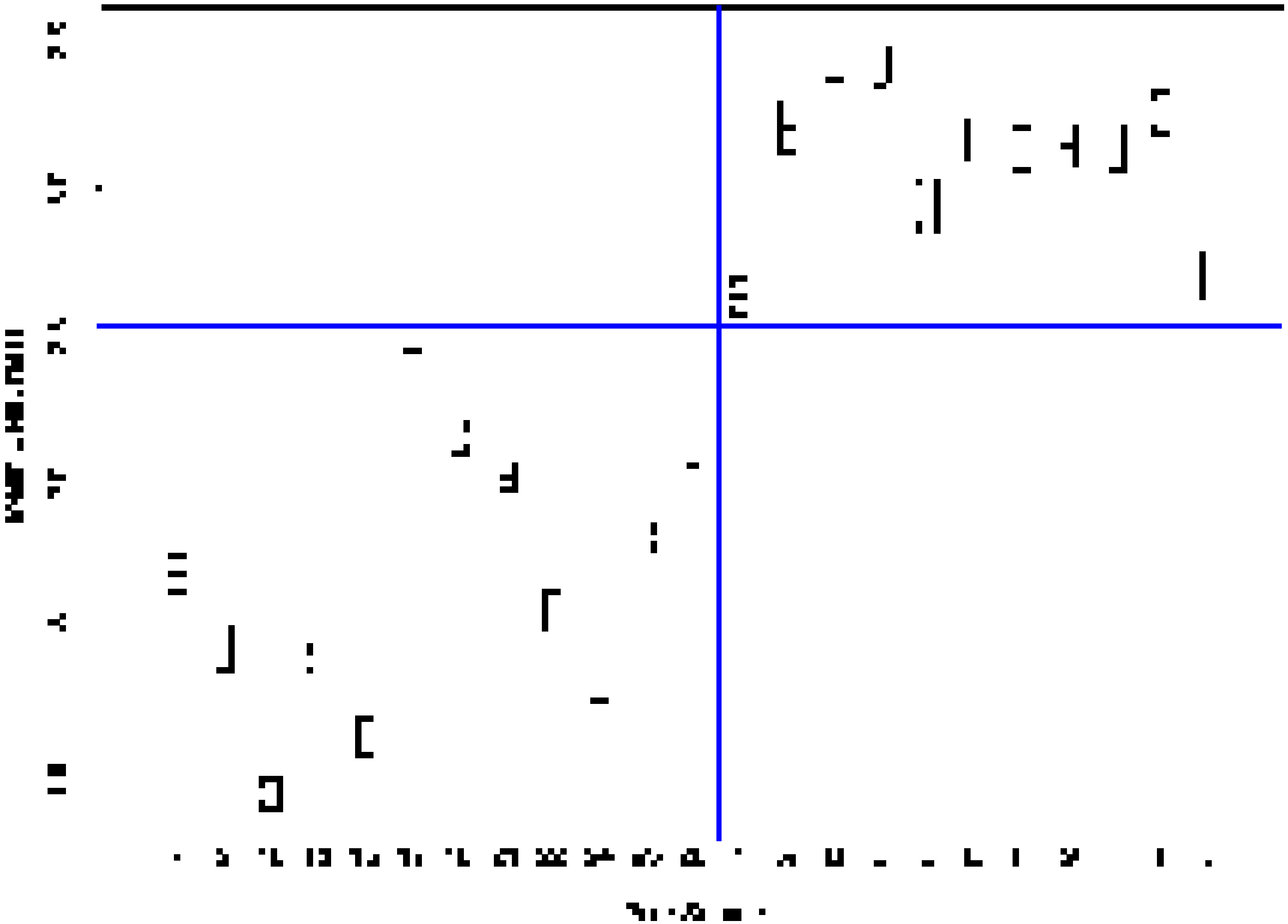
4 U D107.P.10.C.H.H.B.NYT20000415.0068 :: **Public softens on capital punishment; even conservatives questioning fairness, innocence**

1 U D107.P.10.C.H.H.22.NYT20000415.0068 :: **Their views seem incompatible with their political philosophy**

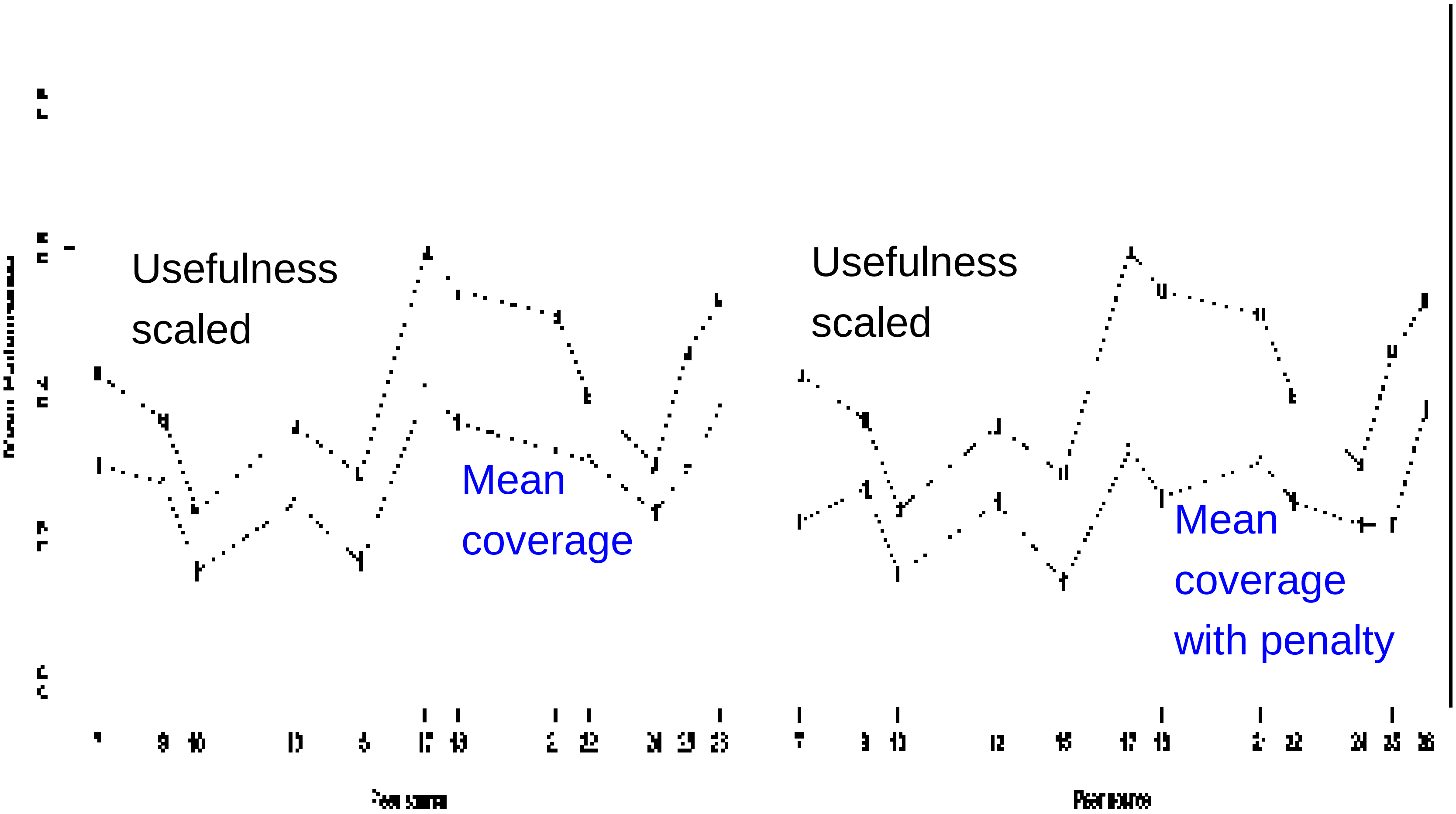
1 U D107.P.10.C.H.H.15.NYT20000415.0068 :: **That people have an incentive to be that the innocent are never to death by state action unborn or in jail whether they are put sure.**

Task 1: Usefulness by peer

~95% confidence intervals around the mean



Task 1: Scaled usefulness & coverage by peer

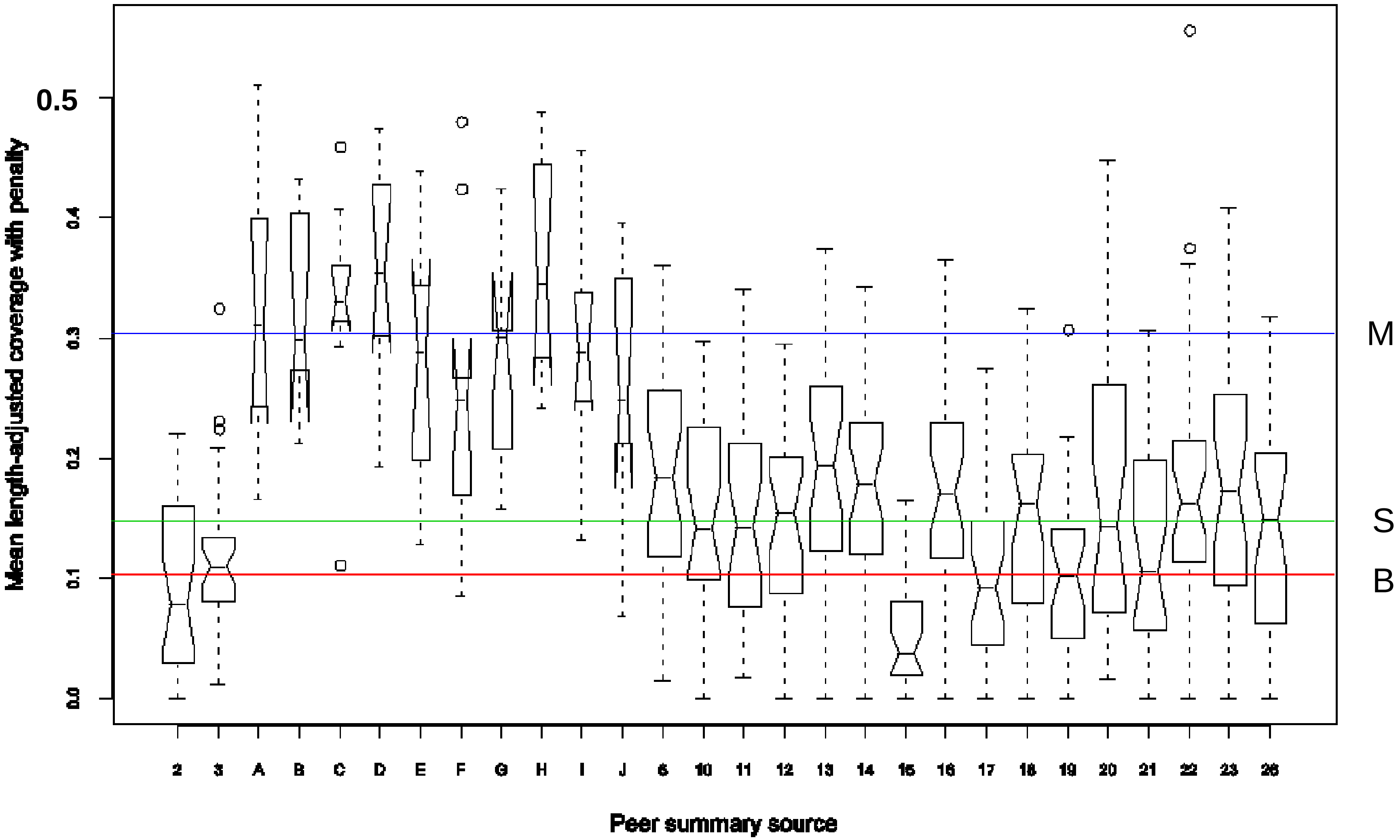


Task 2: Short summary of document set focused by a TDT event topic

- System task:
  - Use the 30 TDT clusters
    - 298 documents
    - ~ 10 documents/cluster
    - ~ 352 sentences/cluster
  - Given:
    - each document cluster
    - the associated TDT topic
  - Create a short summary (~100 words) of the cluster.

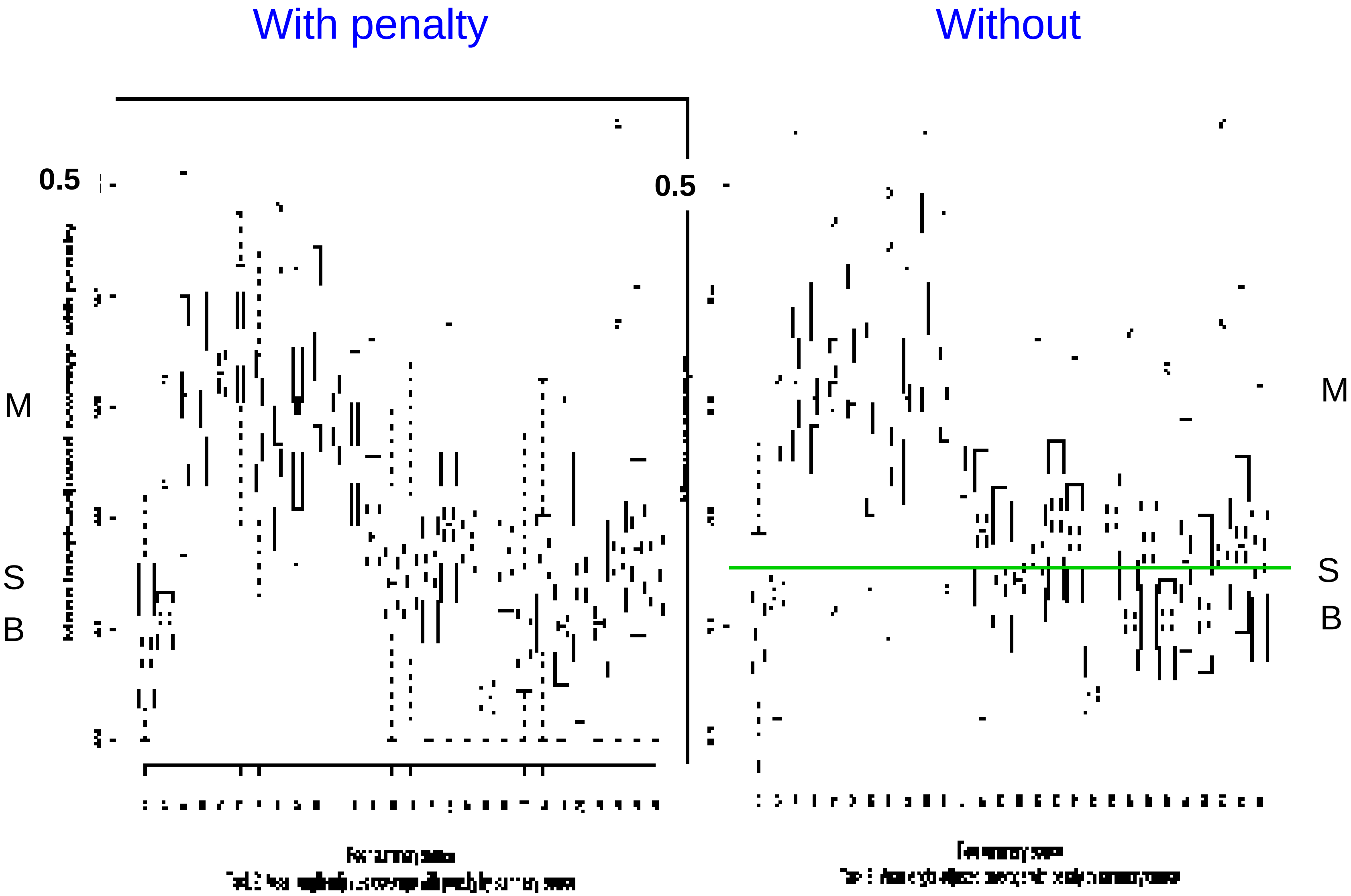
- Evaluation:
  - SEE:
    - 12 linguistic quality items
    - Content coverage
    - Extra material

Task 2: Mean length-adjusted coverage with penalty by peer



Task 2: Mean length-adjusted coverage with penalty by summary source

Task 2: Mean length-adjusted coverage +/- penalty by peer



Tasks 2 - 4: ANOVAs

- Try ANOVA to see if baselines, manual, systems are significantly different from each other as groups.
- ANOVA assumptions/checks:
  - Data approx. normally distributed with approx. equal variances
  - Residuals looked as if they could have come from the same normal distribution
- Results:
  - Task 2: all groups significantly different
    - B != S; S != M; M != B
  - Task 3,4: can't distinguish systems from baselines

	Mean LAC	**Mean LAC, penalty**	**Mean LAC, proportional**
T2			
T3	H^o: B=S	H^o: B=S	H^o: B=S
T4	H^o: B=S	H^o: B=S	H^o: B=S
T4*	H^o: B=S	H^o: B=S	H^o: B=S

* Quadruple judgments

Task 2: Multiple comparisons

```
 REGWQ Grouping      Mean      N    peer
           A      0.18900     30    13
           A
B          A      0.18243     30     6
B          A
B          A      0.17923     30    16
B          A
B          A      0.17787     30    22
B          A
B          A      0.17557     30    23
B          A
B          A      0.17467     30    14
B          A
B          A   C  0.16550     30    20
B          A   C
B   D      A   C  0.15193     30    18
B   D      A   C
B   D      A   C  0.14903     30    11
B   D      A   C
B   D      A   C  0.14520     30    10
B   D      A   C
B   D  E   A   C  0.14357     30    12
B   D  E   A   C
B   D  E   A   C  0.14293     30    26
B   D  E       C
B   D  E       C  0.12583     30    21
    D  E       C
    D  E       C  0.11677     30    3
    D  E
    D  E   F      0.09960     30    19
    D  E   F
    D  E   F      0.09837     30    17
       E   F
       E   F      0.09057     30    2
           F
           F      0.05523     30    15
```

Mean LAC with penalty

```
                    Mean       N   peer
          A       0.32790     30    22
          A
B         A       0.28391     30    13
B         A
B         A       0.27685     30    23
B         A
B         A       0.27465     30    6
B         A
B         A       0.27339     30    16
B         A
B         A       0.27135     30    14
B         A
B         A    C  0.25117     30    20
B         A    C
B    D    A    C  0.23752     30    11
B    D    A    C
B    D    A    C  0.23691     30    18
B    D    A    C
B    D    A    C  0.23628     30    10
B    D         C
B    D    E    C  0.21547     30    12
B    D    E    C
B    D    E    C  0.21422     30    26
B    D    E    C
B    D    E    C  0.18898     30    21
     D    E    C
     D    E    C  0.17561     30    3
     D    E
F    D    E       0.15485     30    19
F    D    E
F    D    E       0.14820     30    17
F         E
F         E       0.13968     30    2
F
F                 0.08211     30    15
```

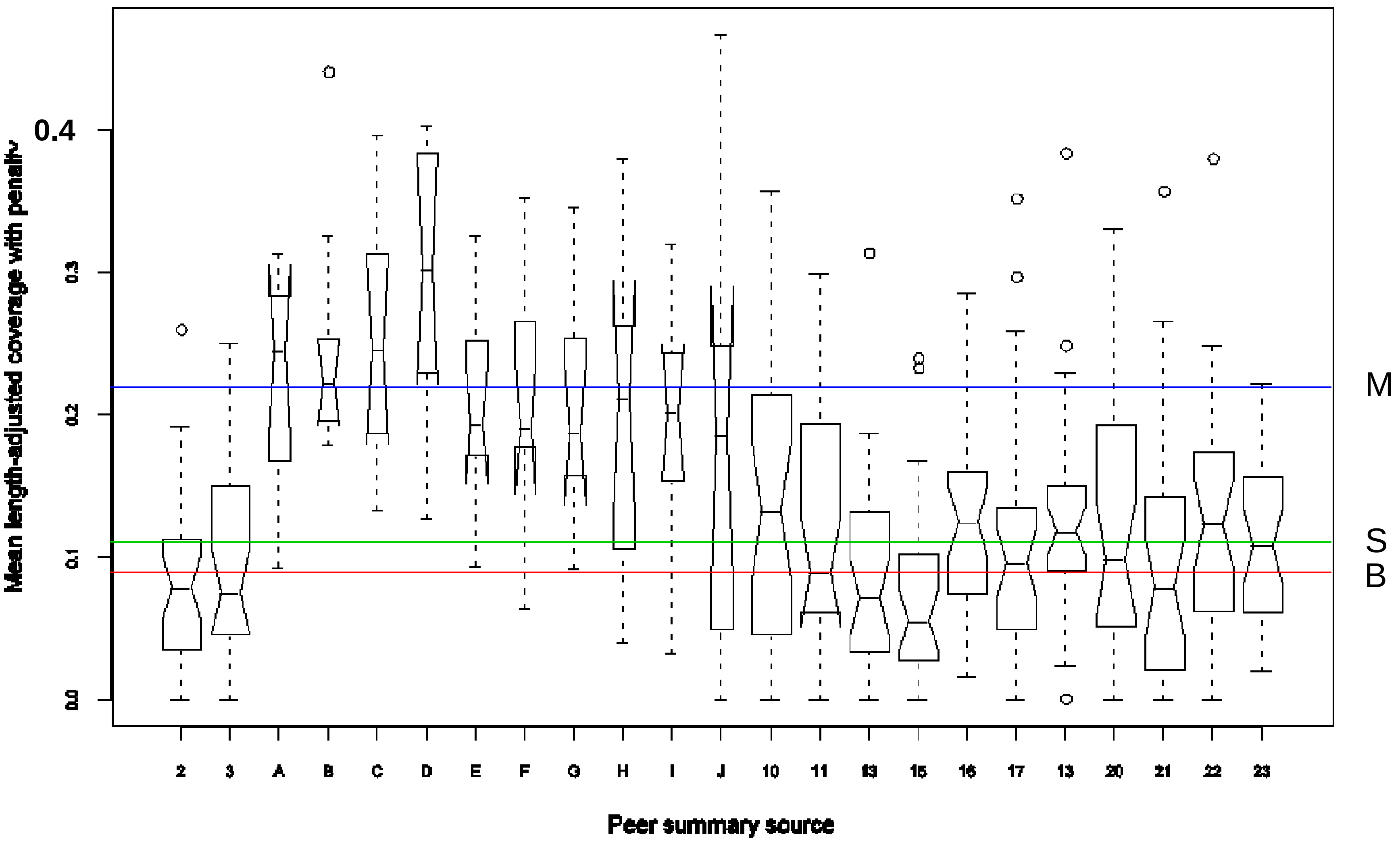
Proportional

Task 3: Short summary of document set focused by a viewpoint statement

- System task:
  - Use the 30 TREC clusters
    - 326 documents
    - ~ 11 documents/cluster
    - ~335 sentences/cluster
  - Given
    - each document cluster
    - a viewpoint description
  - create a short summary (~100 words) of the cluster from the point of view specified.

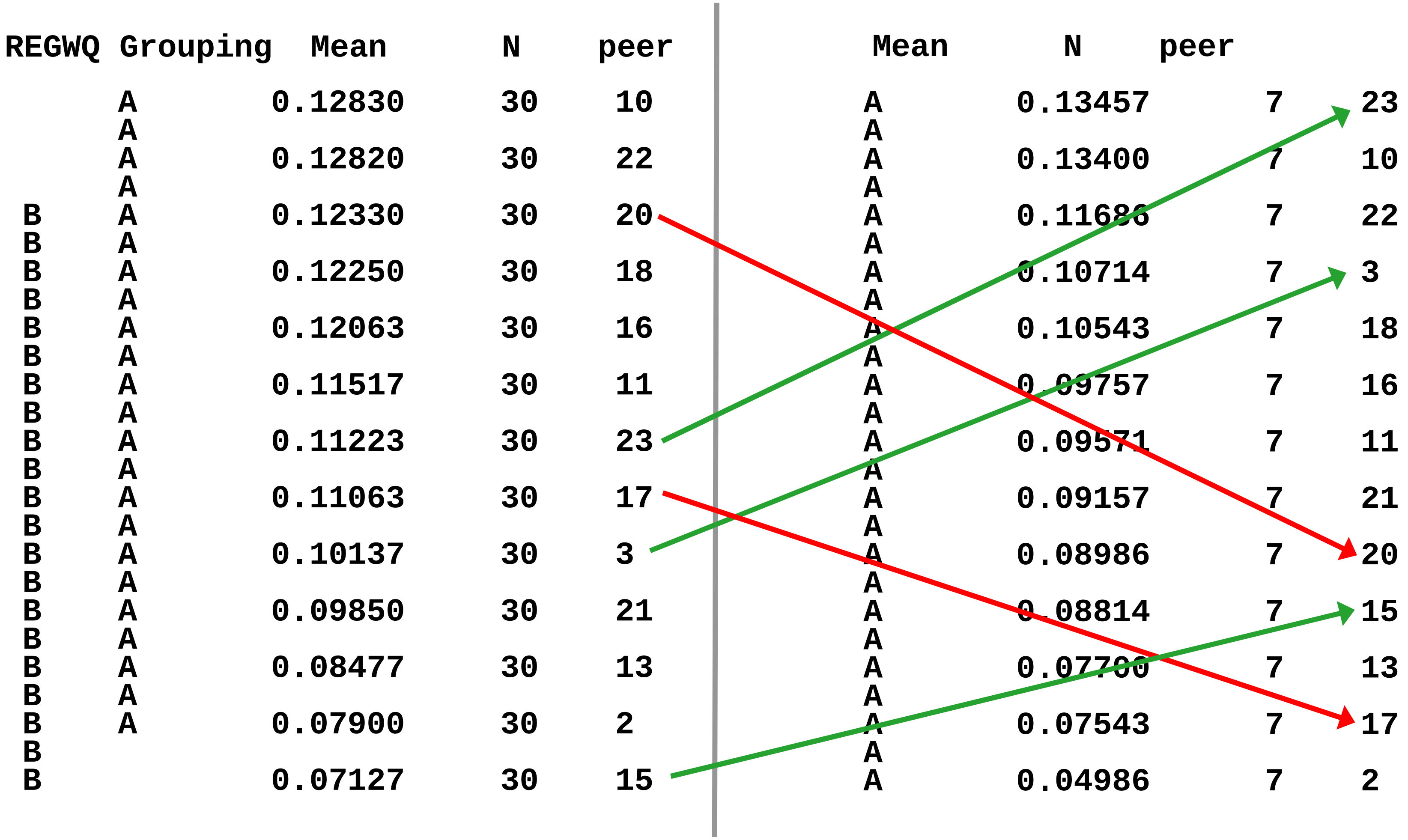
- Evaluation:
  - SEE:
    - 12 linguistic quality items
    - Content coverage
    - Extra material

Task 3: Mean length-adjusted coverage with penalty by peer



Task 3: Mean length-adjusted coverage with penaly by summary source

Task 3: Multiple comparisons



```
REGWQ Grouping   Mean        N     peer

       A       0.12830      30     10
       A
       A       0.12820      30     22
       A
  B    A       0.12330      30     20
  B    A
  B    A       0.12250      30     18
  B    A
  B    A       0.12063      30     16
  B    A
  B    A       0.11517      30     11
  B    A
  B    A       0.11223      30     23
  B    A
  B    A       0.11063      30     17
  B    A
  B    A       0.10137      30     3
  B    A
  B    A       0.09850      30     21
  B    A
  B    A       0.08477      30     13
  B    A
  B    A       0.07900      30     2
  B
  B            0.07127      30     15
```

```
         Mean       N     peer

A         0.13457         7     23
A
A         0.13400         7     10
A
A         0.11686         7     22
A
A         0.10714         7     3
A
A         0.10543         7     18
A
A         0.09757         7     16
A
A         0.09571         7     11
A
A         0.09157         7     21
A
A         0.08986         7     20
A
A         0.08814         7     15
A
A         0.07700         7     13
A
A         0.07543         7     17
A
A         0.04986         7     2
```

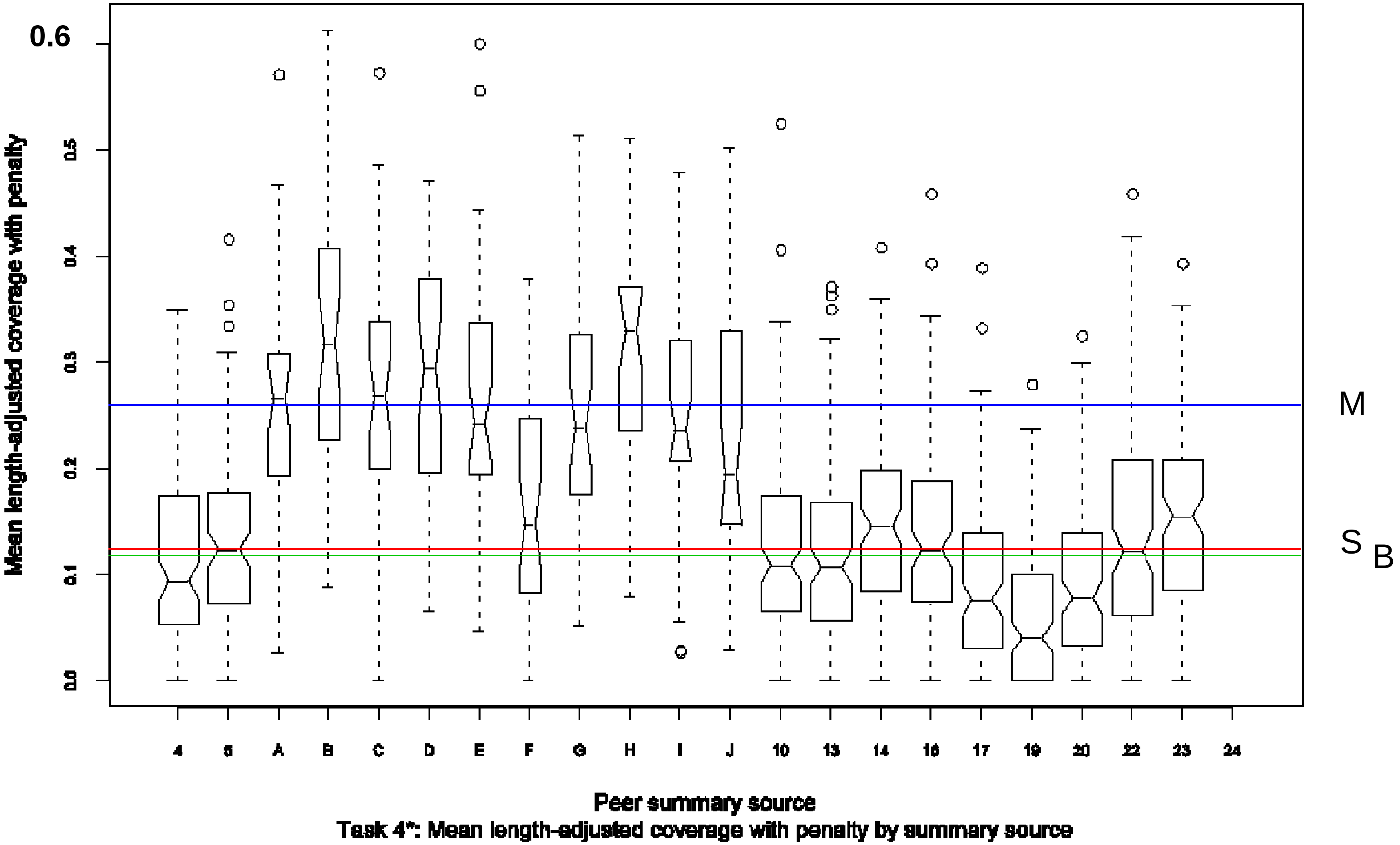
Mean LAC with penalty (full set)   Mean LAC with penalty (subset)

Task 4: Short summary of document set focused by a question

- System task:
  - Use the 30 TREC Novelty track clusters
    - 734 documents
    - ~ 24 documents/cluster
    - ~ 66 relevant sentences/cluster
  - Given:
    - A document cluster
    - A question/topic
    - Set of sentences in each document that are relevant to the question
  - Create a short summary (~100 words) of the cluster that answers the question. Assessors were told to summarize the relevant sentences

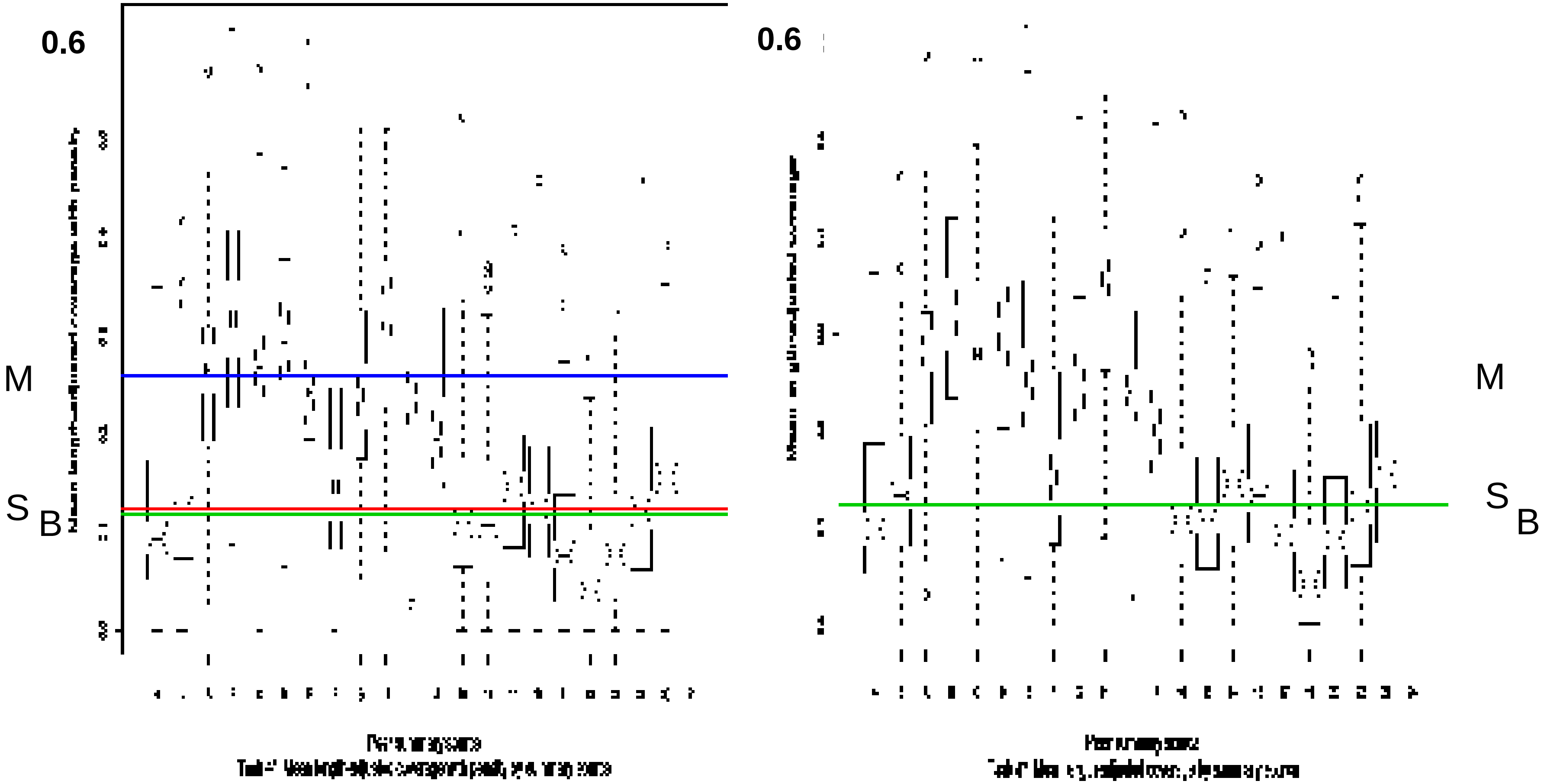
- Evaluation:
  - SEE:
    - 12 linguistic quality items
    - Content coverage
    - Extra material
  - Responsiveness

Task 4*: Mean length-adjusted coverage with penalty by peer



DUC 2003

Task 4*: Mean length-adjusted coverage +/- penalty by peer



With penalty

Without

Task 4*: ANOVA

- Try ANOVA to see if baselines, manual, systems are significantly different from each other *as groups*

- Use quadruple judgment data to estimate effect of interactions

- Model: coverage =

  grandmean +
  docset +
  peer +
  assessor +
  assessorXpeer +
  docsetXpeer +
  docsetXassessor +
  everything else

Task 4*: ANOVA

S-Plus: GLM Procedure using mean LAC, with penalty

```
                                    Sum of
Source                   DF        Squares   Mean Square  F Value
Model                   506     7.87800677    0.01556918     5.37
Error                   787     2.28168160    0.00289921
Corrected Total        1293    10.15968837

       R-Square     Coeff Var      Root MSE      Mean
       0.775418      45.15147      0.053844      0.119253

Source                   DF      Type I SS   Mean Square  F Value   Pr > F
docset                   29     1.31306346    0.04527805    15.62   <.0001
peer                     10     0.94199161    0.09419916    32.49   <.0001
assess                    9     1.24354441    0.13817160    47.66   <.0001
assess*peer              90     0.28209045    0.00313434     1.08   0.2939
docset*peer             289     3.08713511    0.01068213     3.68   <.0001
docset*assess            79     1.01018173    0.01278711     4.41   <.0001
```

Task 4*: Multiple comparisons

```
REGWQ Grouping          Mean        N      peer

         A            0.155814     118      23
         A
         A            0.144517     118      14
    B    A
    B    A      C     0.141136     118      22
    B           C
    B    D      C     0.134596     114      16
    B    D      C
    B    D      C     0.131220     118      5
    B    D      C
    B    D      C     0.123449     118      10
         D      C
         D      C     0.122186     118      13
         D
         D            0.116576     118      4

         E            0.092966     118      17
         E
         E            0.091059     118      20

         F            0.058780     118      19
```

Mean LAC with penalty

```
                      Mean        N      peer

         A          0.24531      118      23
         A
    B    A          0.22017      118      14
    B    A
    B    A     C    0.21548      118      22
    B          C
    B          C    0.20639      118      4
    B          C
    B          C    0.20574      118      10
    B          C
    B          C    0.20327      114      16
    B          C
    B          C    0.19764      118      5
               C
               C    0.18356      118      13

         D          0.14008      118      17
         D
         D          0.13724      118      20

         E          0.09011      118      19
```

Proportional

Means with the same letter are not significantly different.

Task 4: Responsiveness

- Simulated extrinsic evaluation

- Assessor sees
  - the topic for the docset
  - the file of relevant/novel sentences from the docset
  - all summaries of that docset

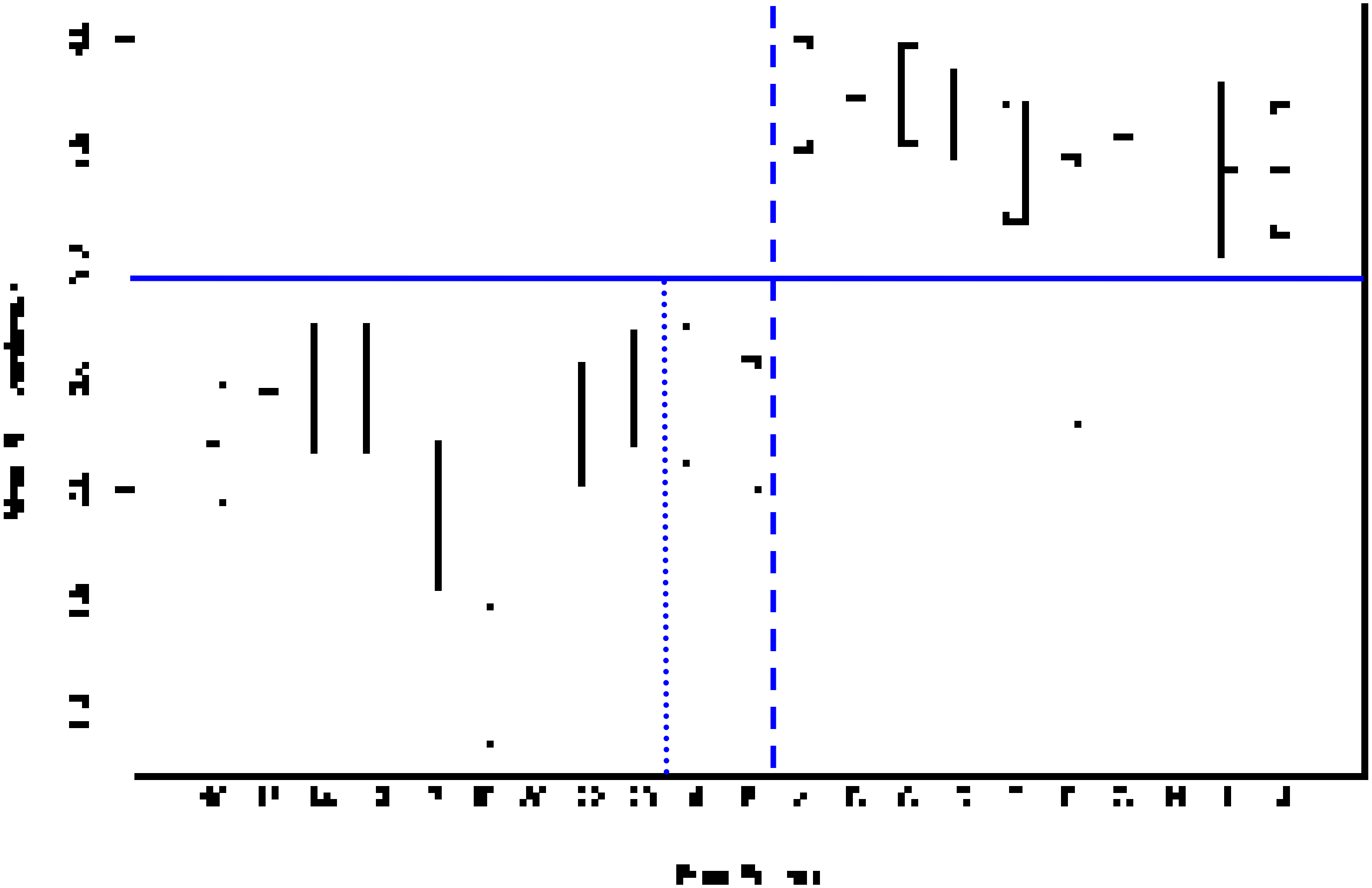
- Assessor asked to:
  - “Read the topic/question and all the summaries.”
  - “Consult the relevant sentences as needed.”
  - “Grade each summary according to how responsive it is in form and content to the question:

    0 (worst), 1, 2, 3, or 4 (best).”

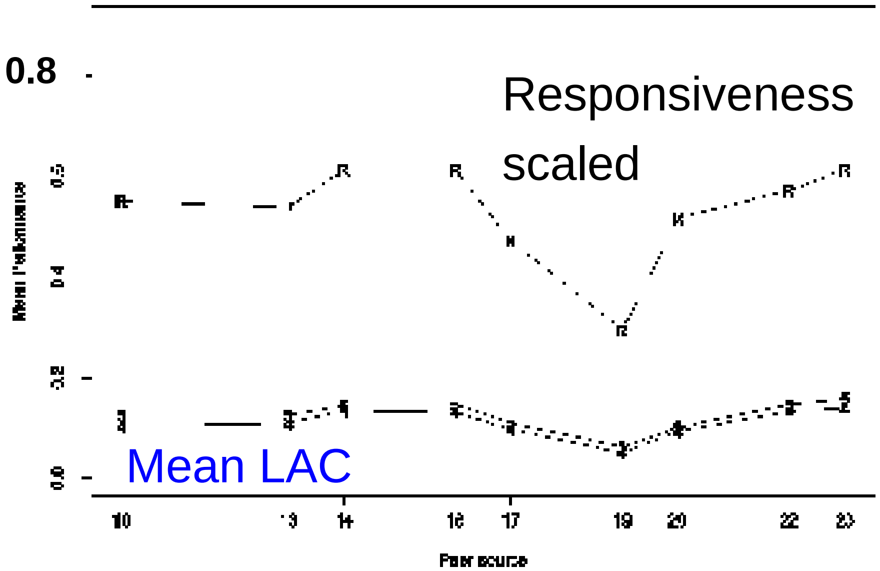
- Double assessment

Task 4: Responsiveness by peer

~95% confidence intervals around the mean

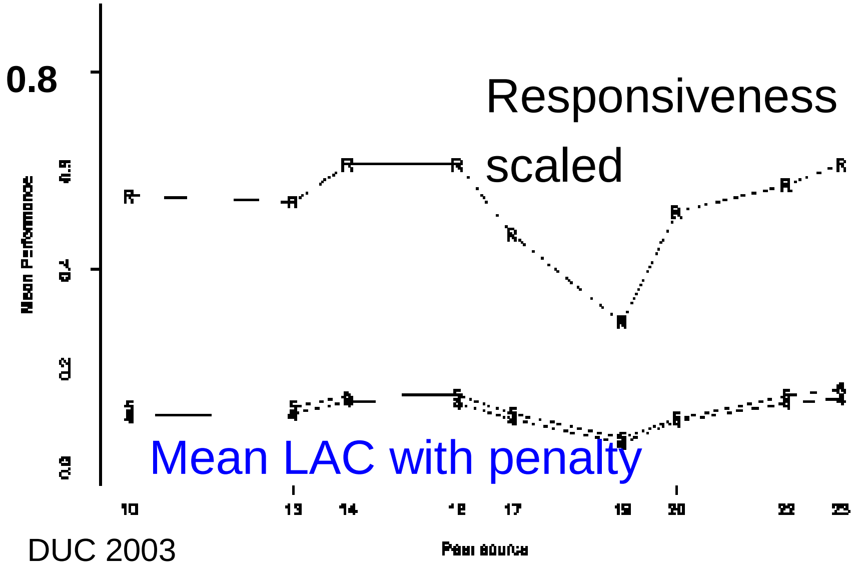


Task 4: Scaled responsiveness vs coverage by peer



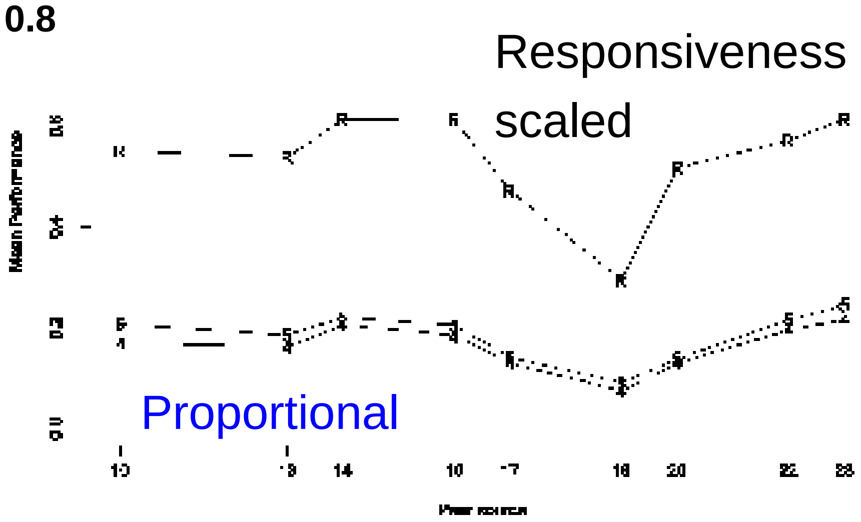
Responsiveness scaled

Mean LAC



Responsiveness scaled

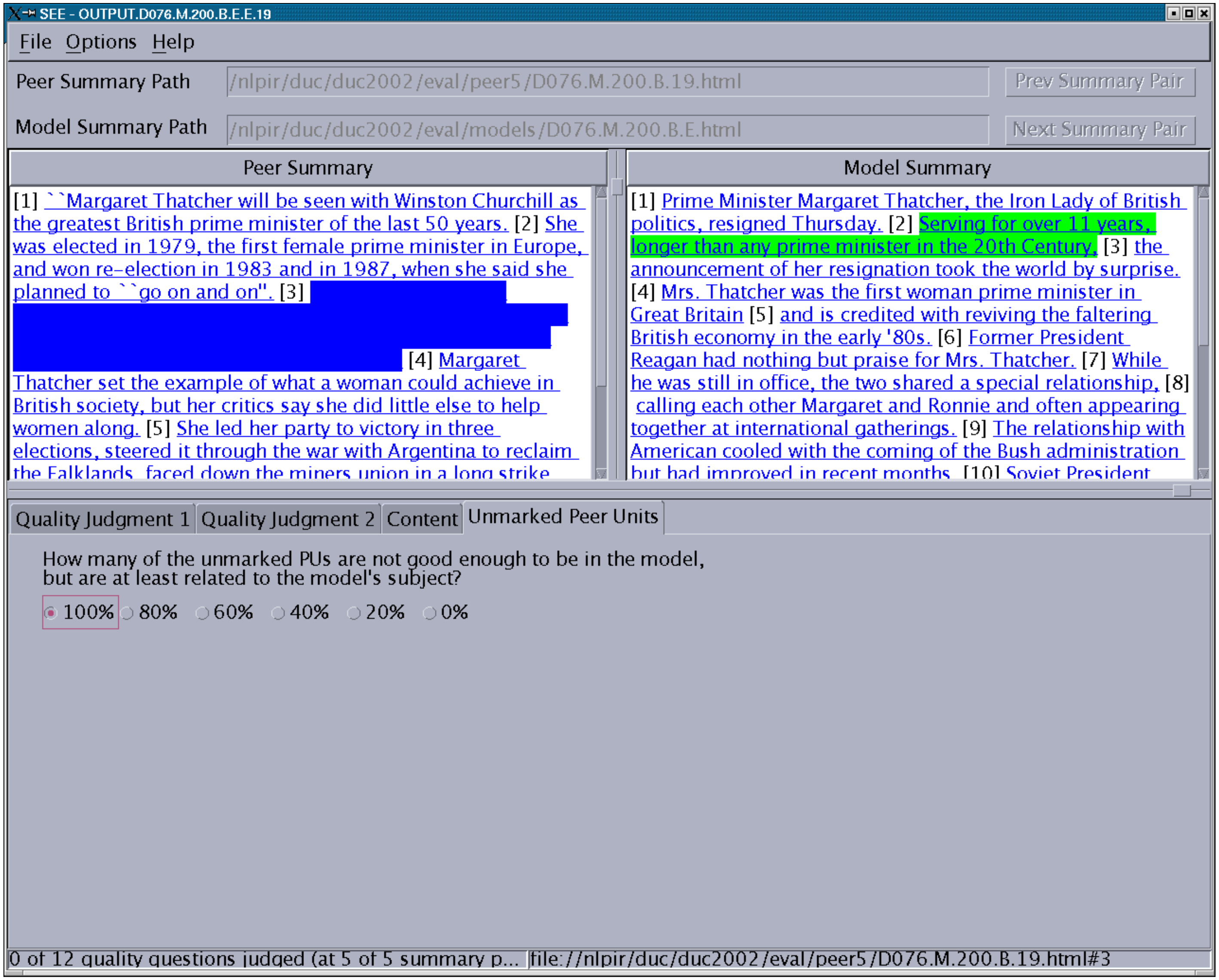
Mean LAC with penalty



Responsiveness scaled

Proportional

SEE: unmarked peer units



Peer Summary Path /nlpir/duc/duc2002/eval/peer5/D076.M.200.B.19.html

Model Summary Path /nlpir/duc/duc2002/eval/models/D076.M.200.B.E.html

Peer Summary

[1] ``Margaret Thatcher will be seen with Winston Churchill as the greatest British prime minister of the last 50 years. [2] She was elected in 1979, the first female prime minister in Europe, and won re-election in 1983 and in 1987, when she said she planned to ``go on and on''. [3] [4] Margaret Thatcher set the example of what a woman could achieve in British society, but her critics say she did little else to help women along. [5] She led her party to victory in three elections, steered it through the war with Argentina to reclaim the Falklands, faced down the miners union in a long strike

Model Summary

[1] Prime Minister Margaret Thatcher, the Iron Lady of British politics, resigned Thursday. [2] Serving for over 11 years, longer than any prime minister in the 20th Century. [3] the announcement of her resignation took the world by surprise. [4] Mrs. Thatcher was the first woman prime minister in Great Britain [5] and is credited with reviving the faltering British economy in the early '80s. [6] Former President Reagan had nothing but praise for Mrs. Thatcher. [7] While he was still in office, the two shared a special relationship, [8] calling each other Margaret and Ronnie and often appearing together at international gatherings. [9] The relationship with American cooled with the coming of the Bush administration but had improved in recent months. [10] Soviet President

How many of the unmarked PUs are not good enough to be in the model, but are at least related to the model's subject?

● 100% ○ 80% ○ 60% ○ 40% ○ 20% ○ 0%

0 of 12 quality questions judged (at 5 of 5 summary p... file://nlpir/duc/duc2002/eval/peer5/D076.M.200.B.19.html#3

Unmarked peer units: evaluation details

- How many of the unmarked peer units are not good enough to be in the model, but at least relevant to the model's subject?

  0% 20% 40% 60% 80% 100% ?

- If the number of unmarked PUs is

  2, choose 0, or 100%

  3, choose 0, 60, or 100%

  4, choose 0, 20, 60, 80, or 100%

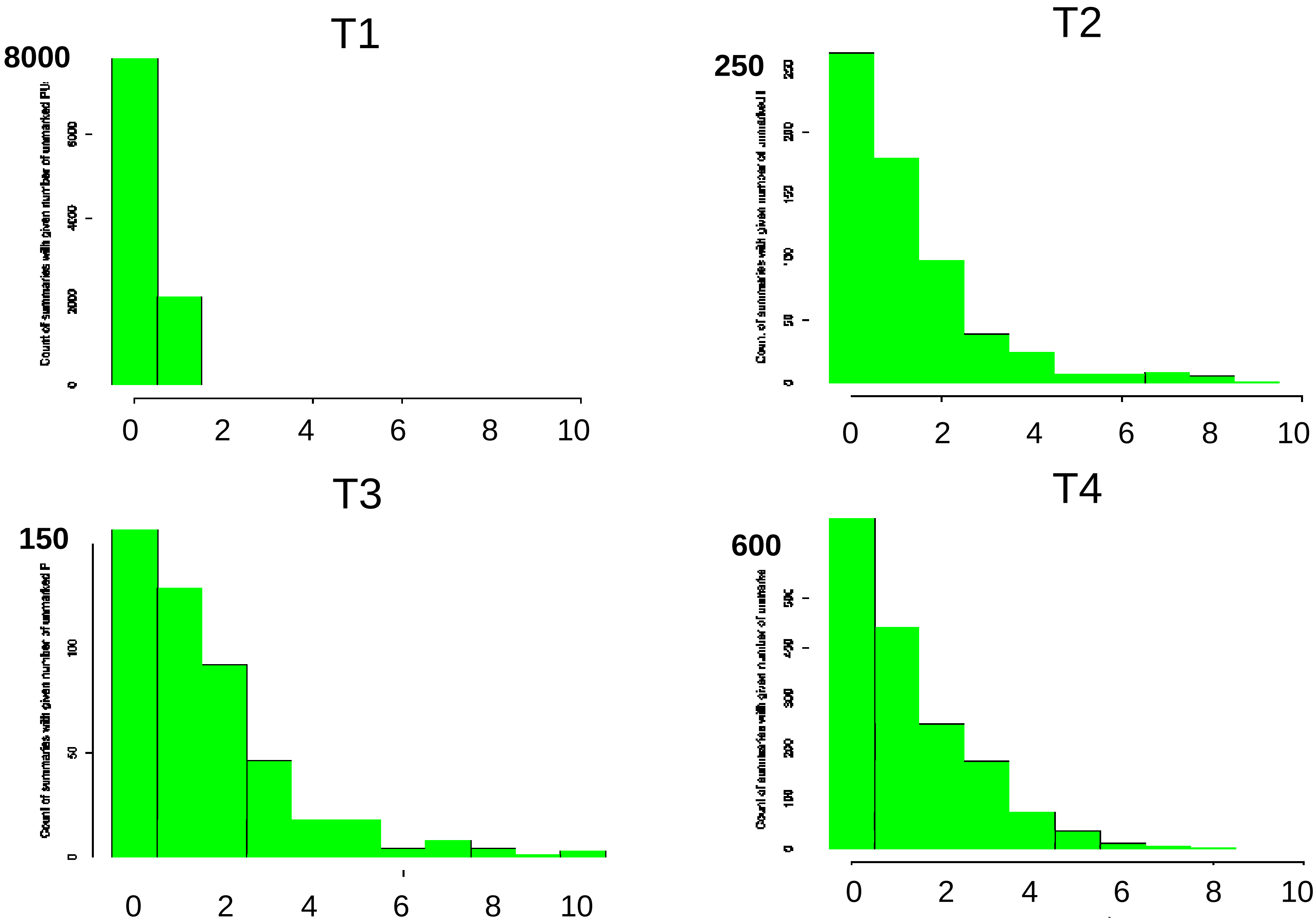
- If half the unmarked PUs are relevant

  Choose 60%

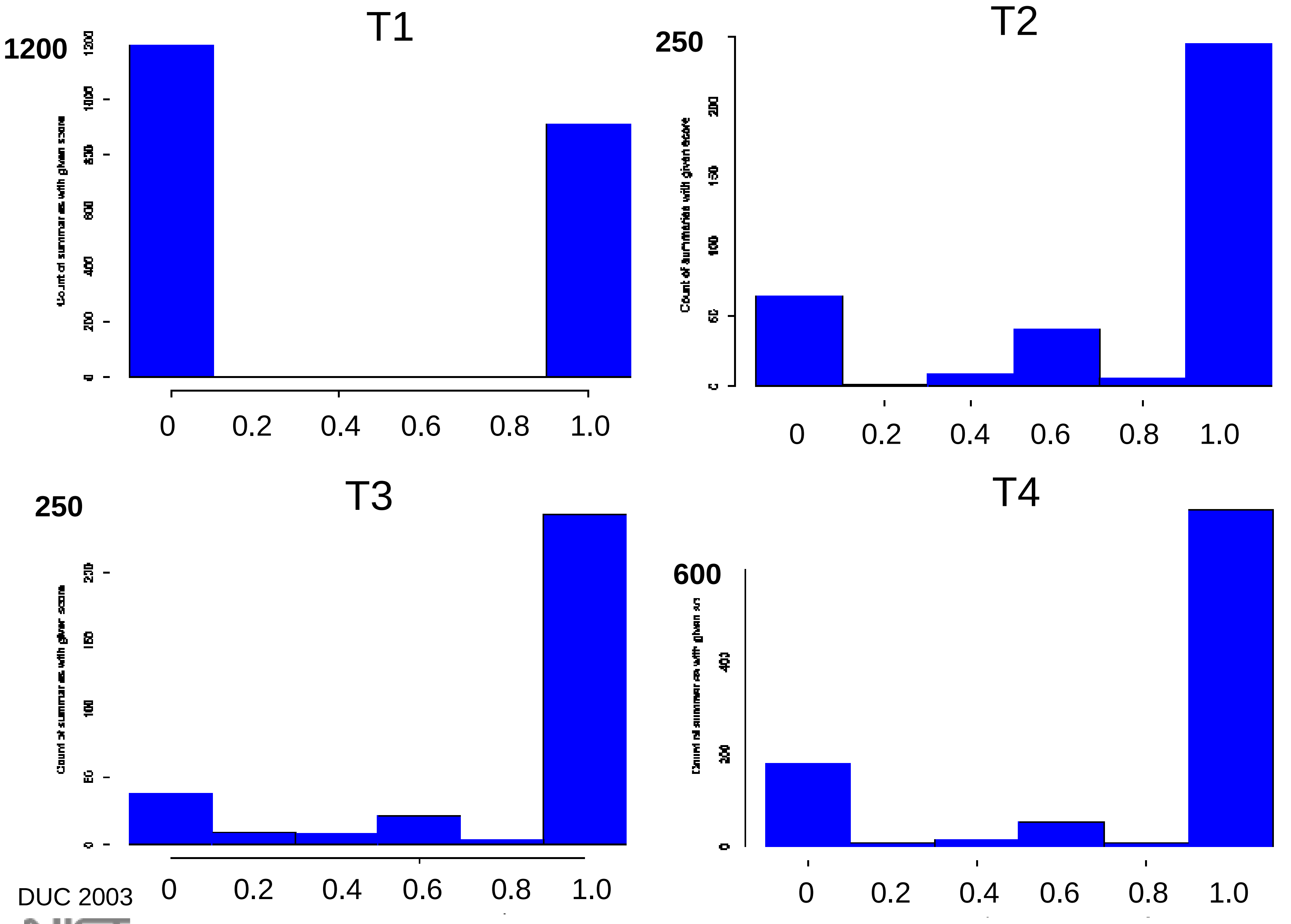
- Mean number of units per summary:

	PUs/summary	MUs/summary
T1	1	1
T2	4.0	10.2
T3	4.1	10.3
T4	3.8	8.8

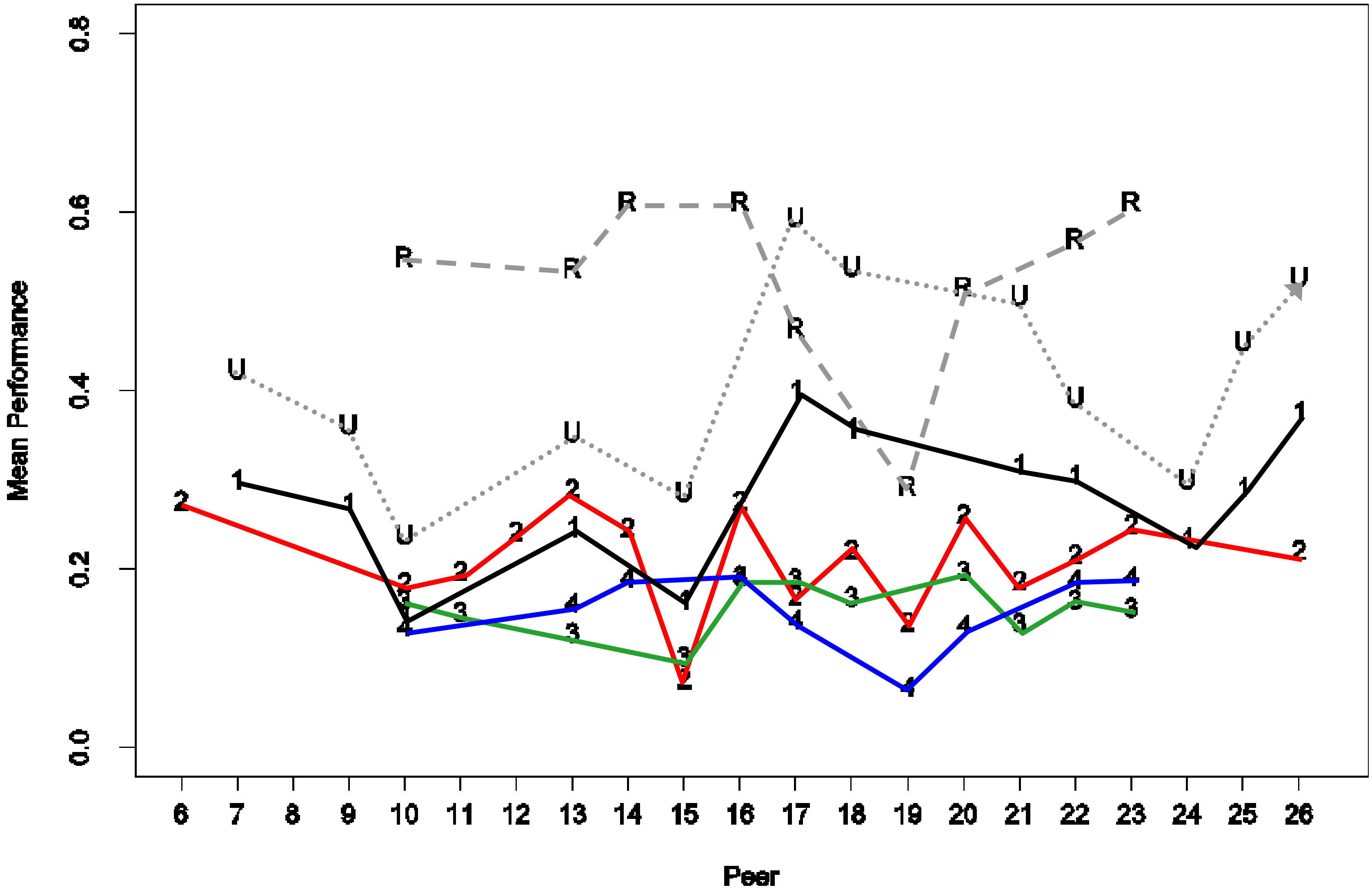
How many abstracts with N unmarked peer units?



How many peer summaries had what % of their unmarked peer units related to model's subject?



Summing up…
Overview of tasks by peer



Summing up …

- Per-unit content (coverage):
  - Still considerable room for system improvement despite large disagreement among humans
  - Most systems indistinguishable from each other in terms of the measures:
    - Task 1
      - Can distinguish a top and a bottom group but not most systems, which are in the middle
    - Task 2,3,4
      - Can distinguish only the systems at each extreme (tasks 2,3) or perhaps bottom group from the rest (task 4)
      - Cannot distinguish systems as a group from baselines in tasks 3,4

Summing up …

- Overall peer quality:
  - Results pass several sanity checks
  - Systems, baselines, and manual are distinguishable
  - Are the “error” conditions too rare to be useful (for largely extractive approaches?)
- Usefulness
  - Manual summaries distinct from systems
  - Tracks coverage for very short summaries
  - Can/should it replace the detailed SEE coverage judgments?
  - Were the lists of keywords more useful then “headline”?
- Responsiveness
  - Manual summaries distinct from systems/baselines
  - Tracks coverage generally
  - Seems doable, but does it measure something different, useful?