Introduction to DUC-2002:
an Intrinsic Evaluation of Generic News Text Summarization Systems

Paul Over
Retrieval Group
Information Access Division

Walter Liggett
Measurement Process Evaluation Group
Statistical Engineering Division

National Institute of Standards and Technology

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, with
  - progressively more demanding text data
  - both direct (intrinsic) and indirect (extrinsic, task-based) evaluations
  - increasing challenge in tasks
- 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)

DUC-2002 schedule

- 26 Nov Call for participation
- 28 Feb Guidelines complete
- 29 Mar Test documents distributed
- 12 Apr Extended abstracts due for speakers
- 15 Apr Results submitted for evaluation
- 7 Jun Evaluated results returned to participants
- 23 Jun Notebook papers due
- 11-12 Jul Workshop at ACL'02 in Philadelphia

Goals of the talk

- Provide an overview of the:
  - Data
  - Tasks
  - Evaluation
    - Experience with implementing the evaluation procedure
    - Feedback from NIST assessors
- Introduce the results:
  - Basics of system performance on 12 + 1 + 1 measures
  - Sanity checking the results and measures
  - Exploration of various factors on performance
    - Systems
    - Document sets, Assessors, Target lengths, Document set types
    - Multi- vs Single document

Data: Formation of test document sets

- Each of 10 NIST information analysts chose one set of newswire/paper articles of each of the following types:
  1. A single natural disaster event with documents created within at most a 7-day window
  2. A single event of any type with documents created within at most a 7-day window
  3. Multiple distinct events of the same type (no time limit)
  4. Biographical (discuss a single person)

- Each assessor chose 2 more sets of articles so that we ended up with a total of 15 document sets of each type.

- Each set contains about 10 documents

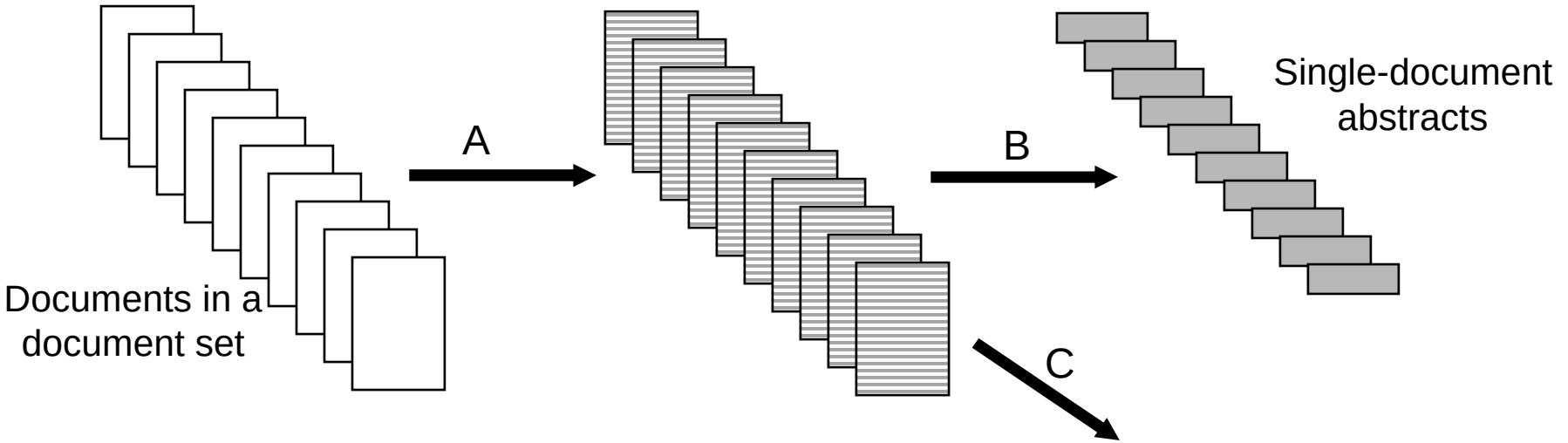
- All documents in a set to be mainly about a specific “concept”

Example document set subjects

- Hurricane Gilbert (1)
- Outcome of longest criminal trial in US history (2)
- Grievances & strikes of miners around the world (3)
- Andrei Sakharov (4)

- The eruption of Mt. Pinatubo in the Philippines (1)
- The Clarence Thomas confirmation hearings (2)
- Heart attacks (3)
- Margaret Thatcher (4)

Manual abstract creation

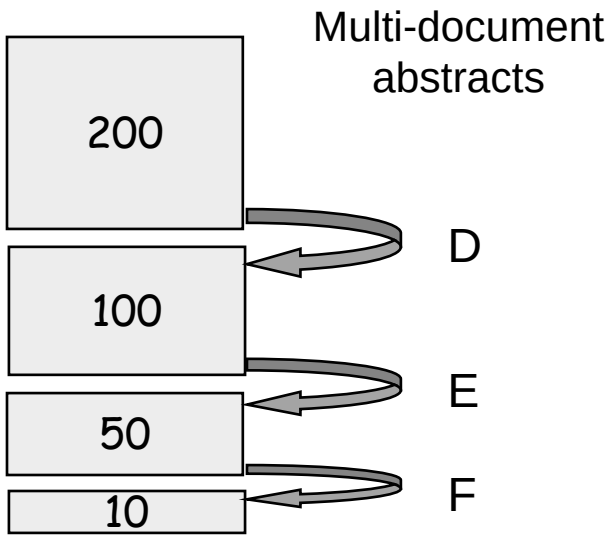


A: **Read hardcopy of documents.**

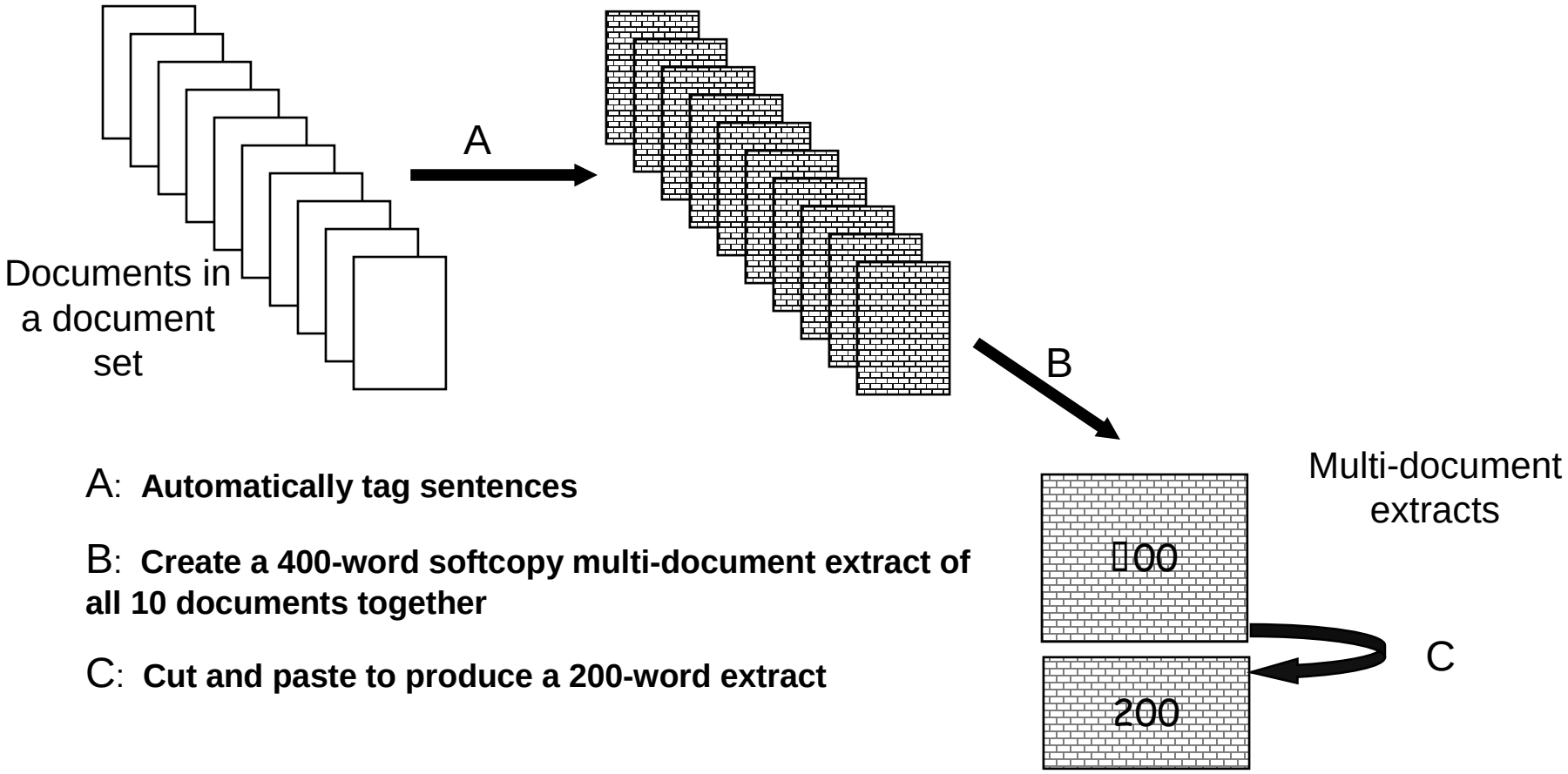
B: **Create a 100-word softcopy abstract for each document using the document author's perspective.**

C: **Create a 200-word softcopy multi-document abstract of all 10 documents together written as a report for a contemporary adult newspaper reader.**

D,E,F: **Cut, paste, and reformulate to reduce the size of the abstract by half.**



Manual extract creation



A: **Automatically tag sentences**

B: **Create a 400-word softcopy multi-document extract of all 10 documents together**

C: **Cut and paste to produce a 200-word extract**

Duplicate and withdrawn abstracts/extracts

- NIST created two sets of abstracts and extracts for each of the 60 document sets

- NIST withdrew – due to differences in documents used by systems and NIST summarizers – the following:

  – D076: one set of abstracts and extracts
  – D088: both sets of abstracts and extracts
  – D098: one set of abstracts and extracts

Automatic baseline abstracts

- NIST (Nega Alemayehu) created 3 baselines automatically based roughly on algorithms suggested by Daniel Marcu from earlier work

- No truncation of sentences, so some baseline abstracts went over the limit (+ <=15 words) and some were shorter than required

- Algorithms:
  1. Single-document summaries:
     - take the first 100 words in the document
  2. Multi-document summaries:
     - take the first 50, 100, or 200 words in the most recent document.
  3. Multi-document summaries:
     - take the first sentence in the 1st, 2nd, 3rd,… document in chronological sequence until you have the target summary size.

Submitted summaries by system code

Abstracts					Extracts				
Single	Multi								
100	10	50	100	200	200	400	System ID	Code	Group
567	0	0	0	0	0	0	uottawa	15	Univ. of Ottawa
567	59	59	59	59	59	59	MICHIGAN	16	Univ. of Michigan
565	0	0	0	0	0	0	SumUMFAR	17	Univ. of Montreal
567	0	0	0	0	0	0	imp_col	18	Imperial College
566	59	59	59	59	59	59	lcc.duc02	19	LCC
0	59	59	59	59	59	59	tno-duc02	20	TNO
567	0	0	0	0	59	59	wpdv-xtr.v1	21	Catholic Univ. Nijmegen
0	0	0	0	0	59	59	unicorp.v36	22	Pennsylvania State Univ.
559	0	0	0	0	0	0	MSRC	23	Microsoft
0	0	59	59	59	59	59	lion_sum	24	Columbia Univ.
566	59	59	59	59	59	59	gleans.v1	25	ISI/Gleans
0	59	59	59	59	0	0	webcl2002	26	ISI/Webclopedia
567	0	0	0	0	0	0	ntt.duc02	27	NTT
567	0	59	59	59	59	59	ccsnsa.v2	28	CCS-NSA
567	59	59	59	59	59	59	kul.2002	29	Catholic Univ. Leuven
567	0	0	0	0	0	0	bbn.headln	30	BBN
567	0	0	0	0	59	59	ULeth131m	31	Univ. of Lethbridge
7359	354	472	472	472	590	590			
9129					1180				

Evaluation basics

- Intrinsic evaluation by humans using special version of SEE (thanks to Chin-Yew Lin, 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

Models

- Source:
  - Authored by a human
  - Phase 1: assessor is model author, but not the document selector
  - Phase 2: assessor is neither document selector nor model author

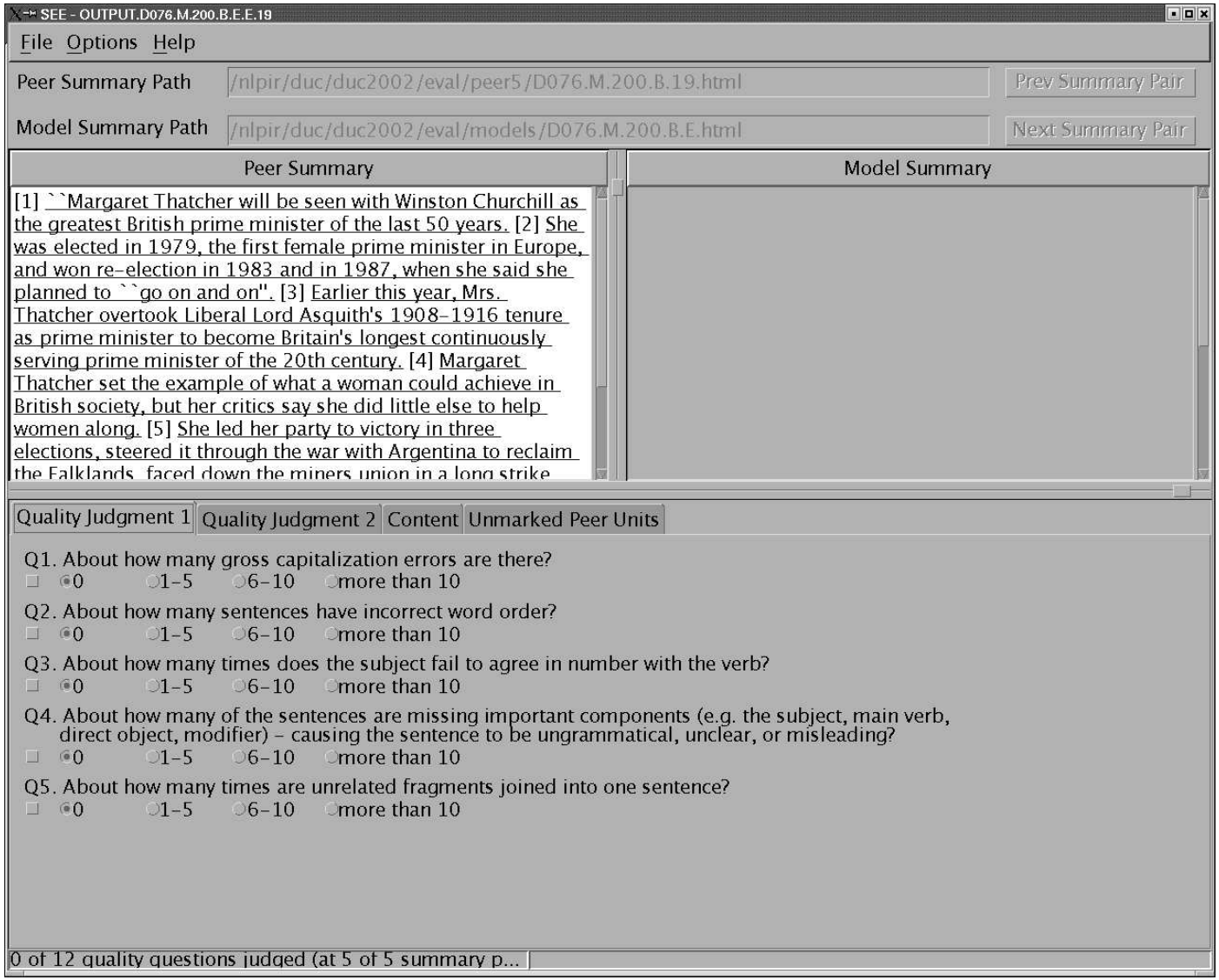
- Formatting:
  - Divided into model units (MUs)
    - (MUs == EDUs - thanks to Alexander Fraser at ISI)
  - Lightly edited by authors to integrate uninterpretable fragments
    - George Bush's selection of Dan Quale
    - as his running mate surprised many
    - many political observers thought him 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
    - Flowed together with HTML tags for SEE

- 3 Sources:
  1. Automatically generated by research systems
     - For single-document summaries: 5 “randomly” selected from those abstracted by all systems
  2. Automatically generated by baseline algorithms
  3. Authored by a human other than the assessor

SEE: overall peer quality



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% Conf Int	Max
Multi-doc:				
Systems	1770	1.821	1.741 – 1.901	9
Baselines	354	1.234	1.094 – 1.374	7
Manuals	228	0.539	0.419 – 0.659	5
Single-doc:				
Systems	3827	1.276	1.236 – 1.316	10
Baseline	294	0.718	0.598 – 0.838	8
Manuals	285	0.505	0.405 – 0.605	5

Overall peer quality

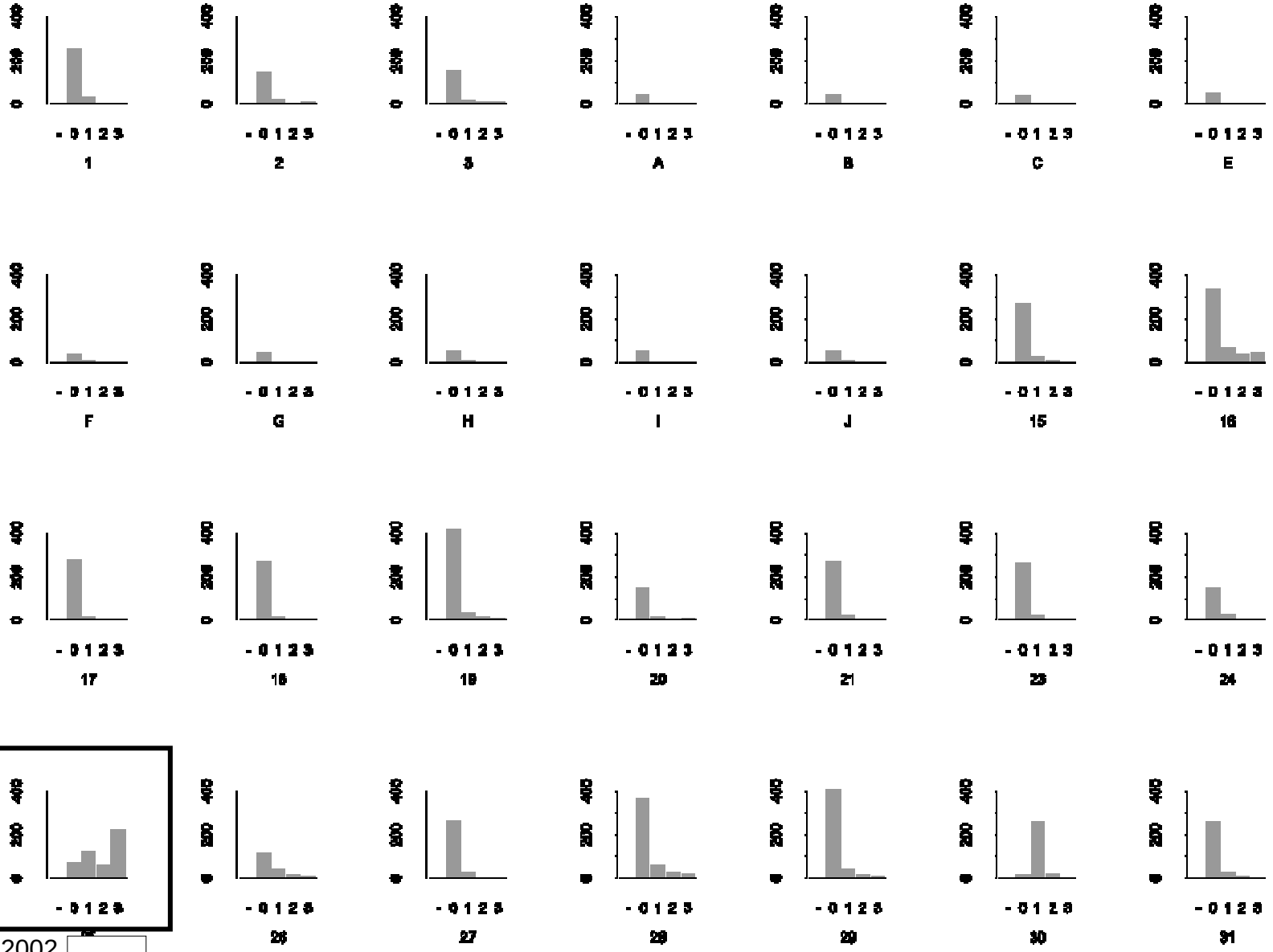
Uneven distribution of scores by question

	None	*1-5*	*6-10*	*> 10*
Q1	4847	904	236	360
Q2	5955	391	1	
Q3	6188	159		
Q4	4932	1408	6	1
Q5	5665	669	8	5
Q6	6084	260	3	
Q7	5796	545	6	
Q8	5092	1245	7	3
Q9	6219	128		
Q10	6164	183		
Q11	5778	557	11	1
Q12	4423	1857	66	1

- Capitalization
- Main component missing
- Unrelated fragments joined
- Noun referent unclear
- Unnecessary repetition
- Misplaced sentences

Overall peer quality

Q1: Capitalization by peer source

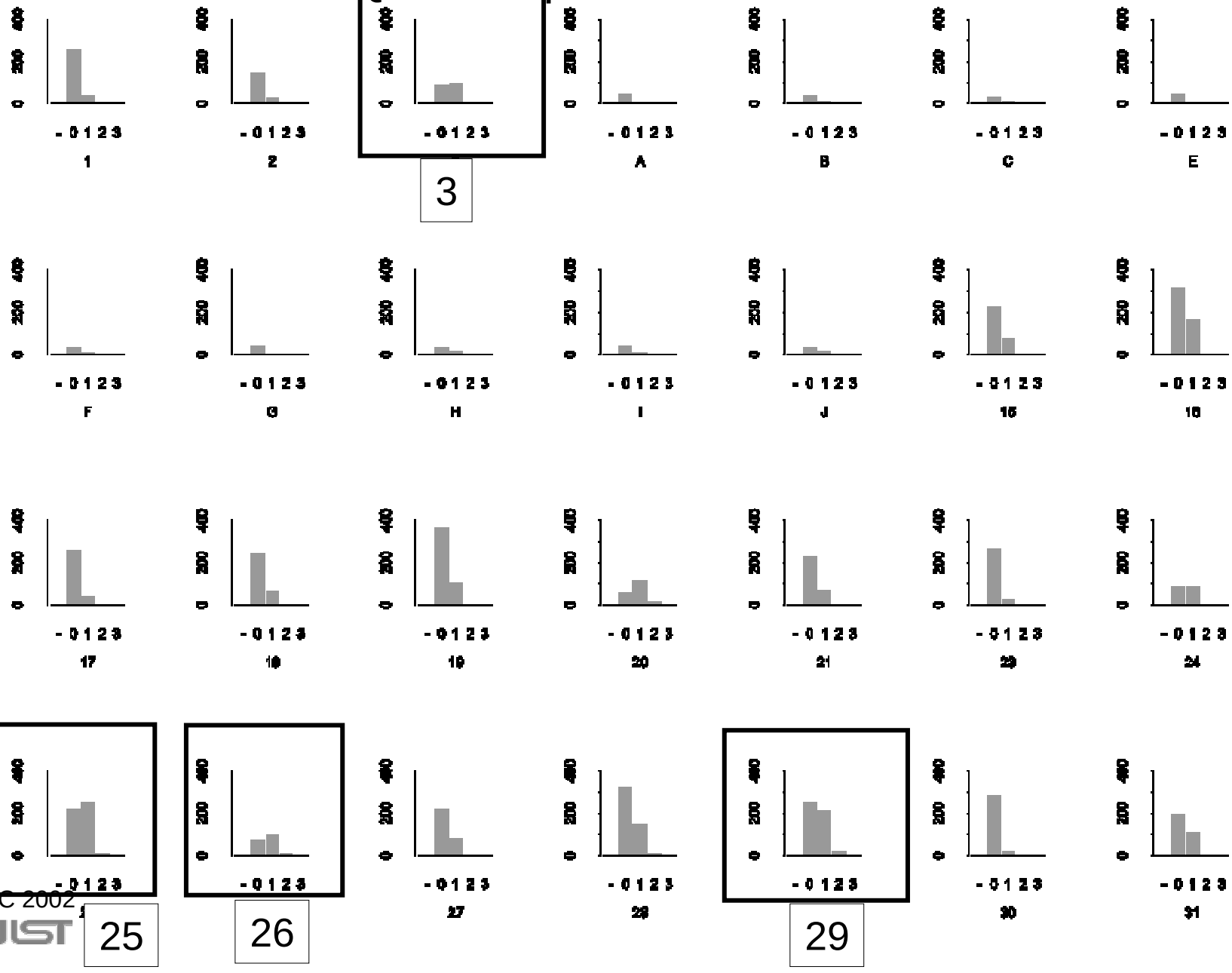


System 25 (GLEANS) added headlines

[1] RIFLES IN LODI, SACRAMENTO, AND OHIO [2] A series of rifles happened in Lodi, Sacramento, Ohio, and other places between Jan. 17, 1989 and Jan. 21, 1989. [3] "Several dozen shots were heard from an automatic rifle," said Monk.

Overall peer quality

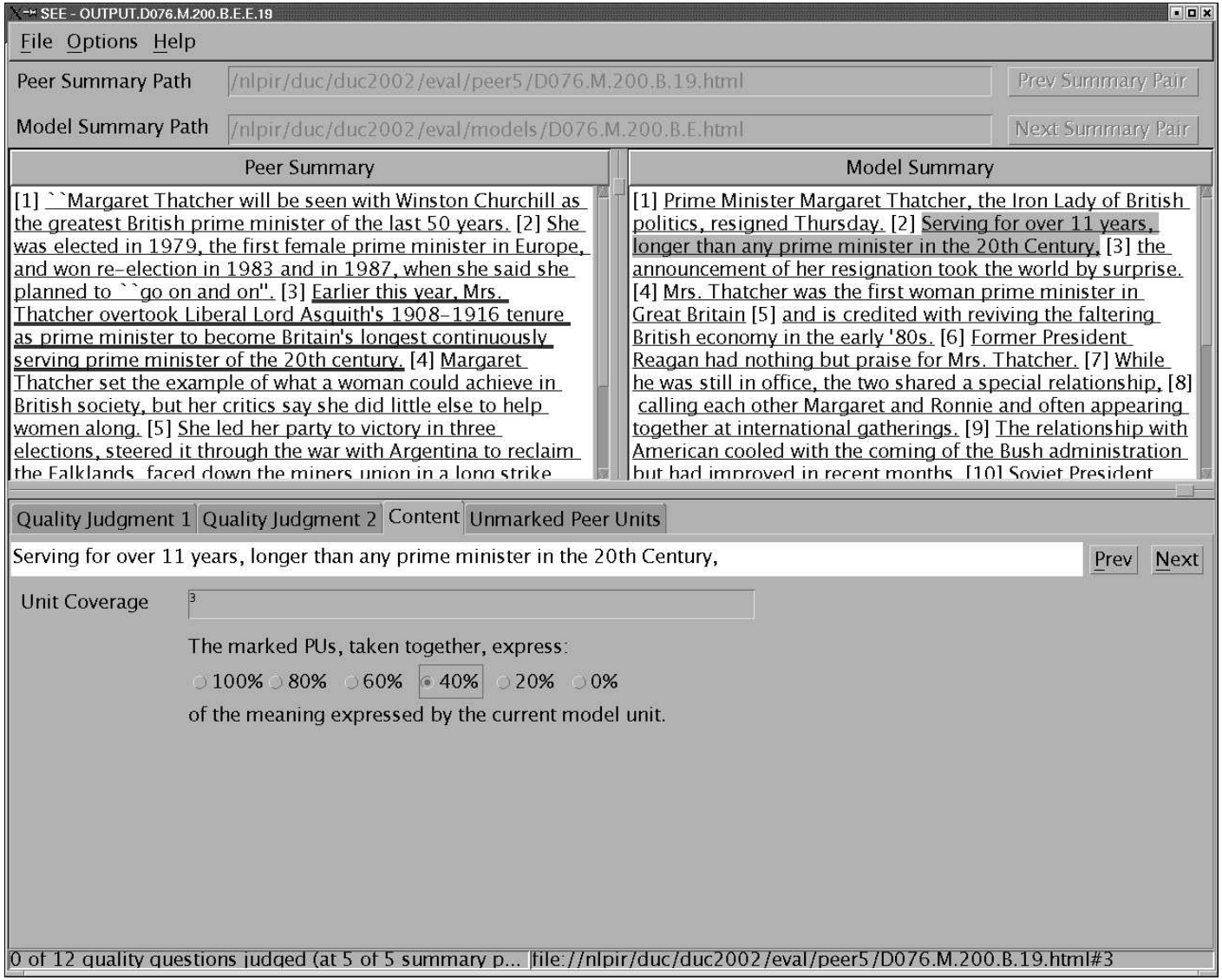
Q12: Misplaced sentences



Overall peer quality
Assessor feedback

- Not sure how to count capitalization errors in “all caps” headline
- Some accepted extra material (place, news service, …); others called it a fragment.
- British versus American English, e.g., “in hospital”
- Sometimes domain knowledge (e.g., place names) made a difference in judging coverage
- Sometimes fragments were related but joined awkwardly – no question to catch this
- Tended to step through text and then look for relevant question rather than step through questions and look for relevant text
- Peer unit boundaries were distracting

SEE: per-unit content



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.

- "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"

- Lots of judgments:
  - 6 742 abstracts judged
  - 63 320 MUs
  - 276 697 MU-PU comparisons

Per-unit content: assessor feedback

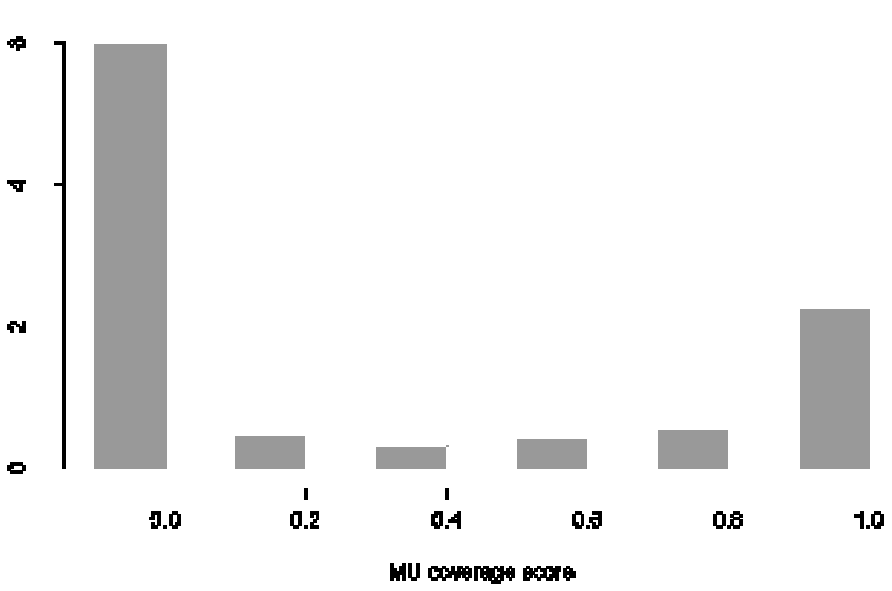
- Missed “50%” choice among the possible answers
- Some confusion about criteria for marking peer units:
  - Share expression of some assertions?
  - Share references to same people, places, things,..?
- Some model units not large enough to express an assertion and so could not overlap with any peer unit.

Per-unit content: measures

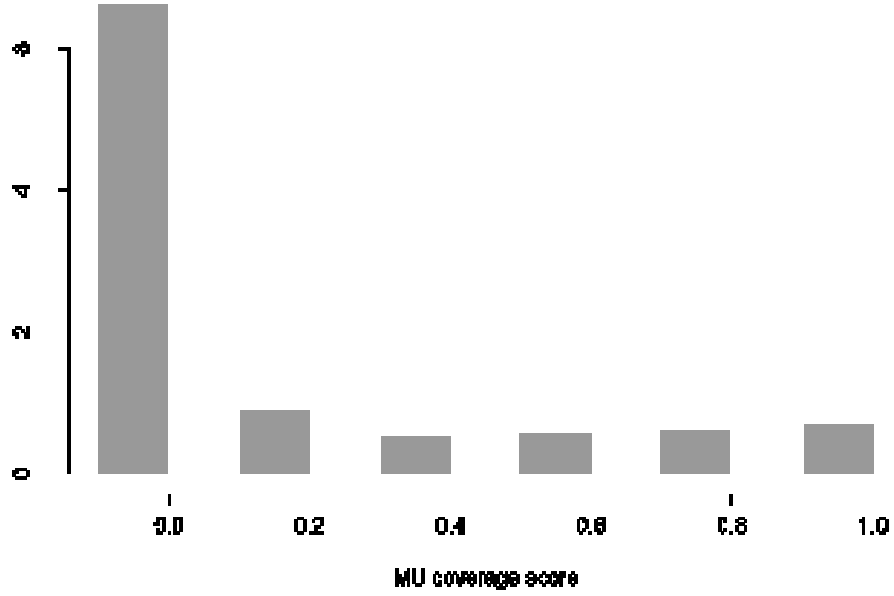
- Recall
  - What fraction of the model content is also expressed by peer?
  - Mean coverage –
    - average of the per-MU completeness judgments [0, 20, 40, 60, 80,100]% for a peer summary
  - Mean length-adjusted coverage –
    - average of the per-MU length-adjusted coverage judgments for a peer
    - 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

Per-unit content: Distribution of individual MU coverage scores

Single-document abstracts



Multi-document abstracts

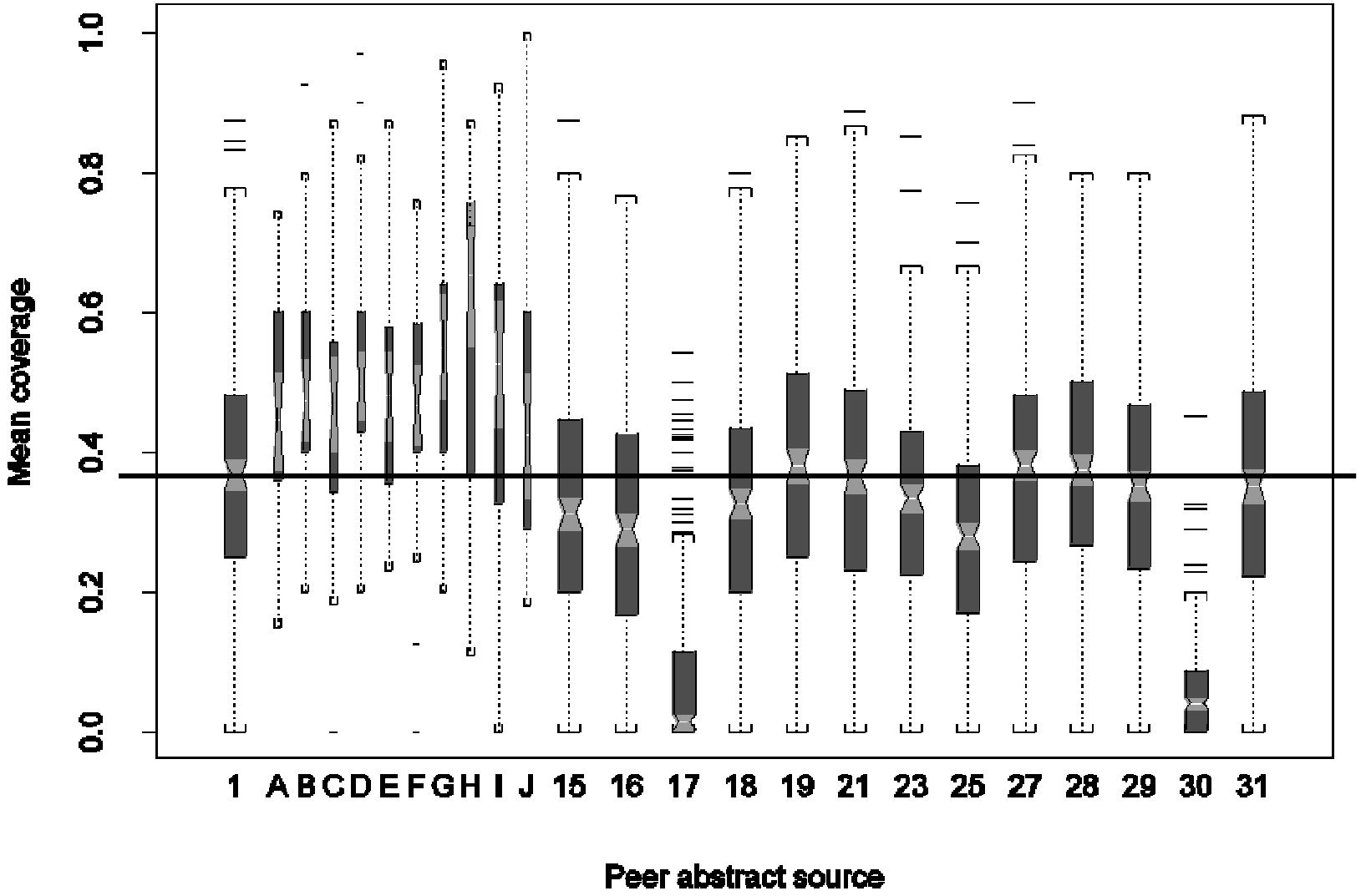


Most MUs (62%) have 0% coverage (~42% for manually created peers)

63% of MUs had no coverage in DUC-2001

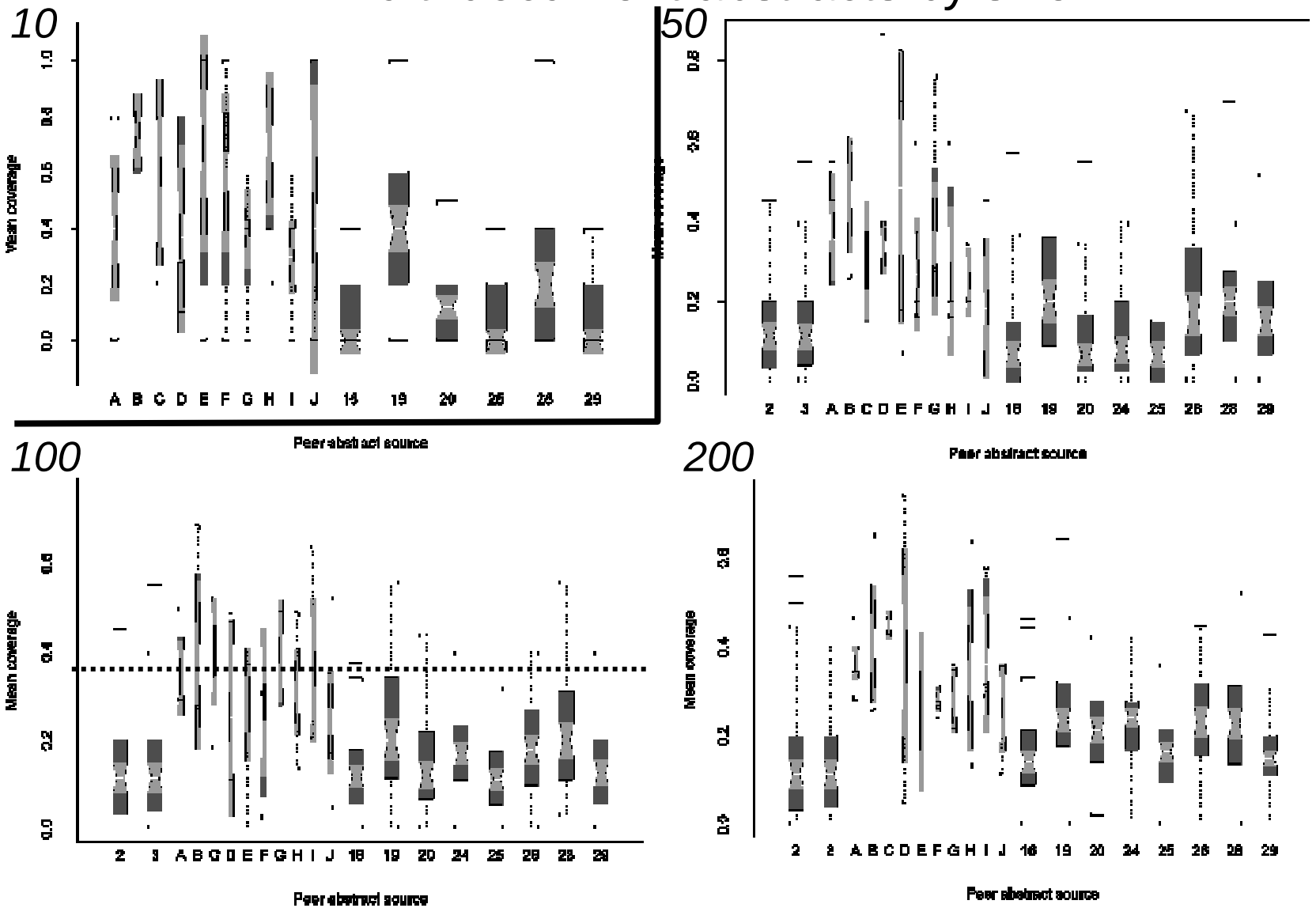
Appears to be due to real differences in content

Mean coverage by peer source
Single-document abstracts

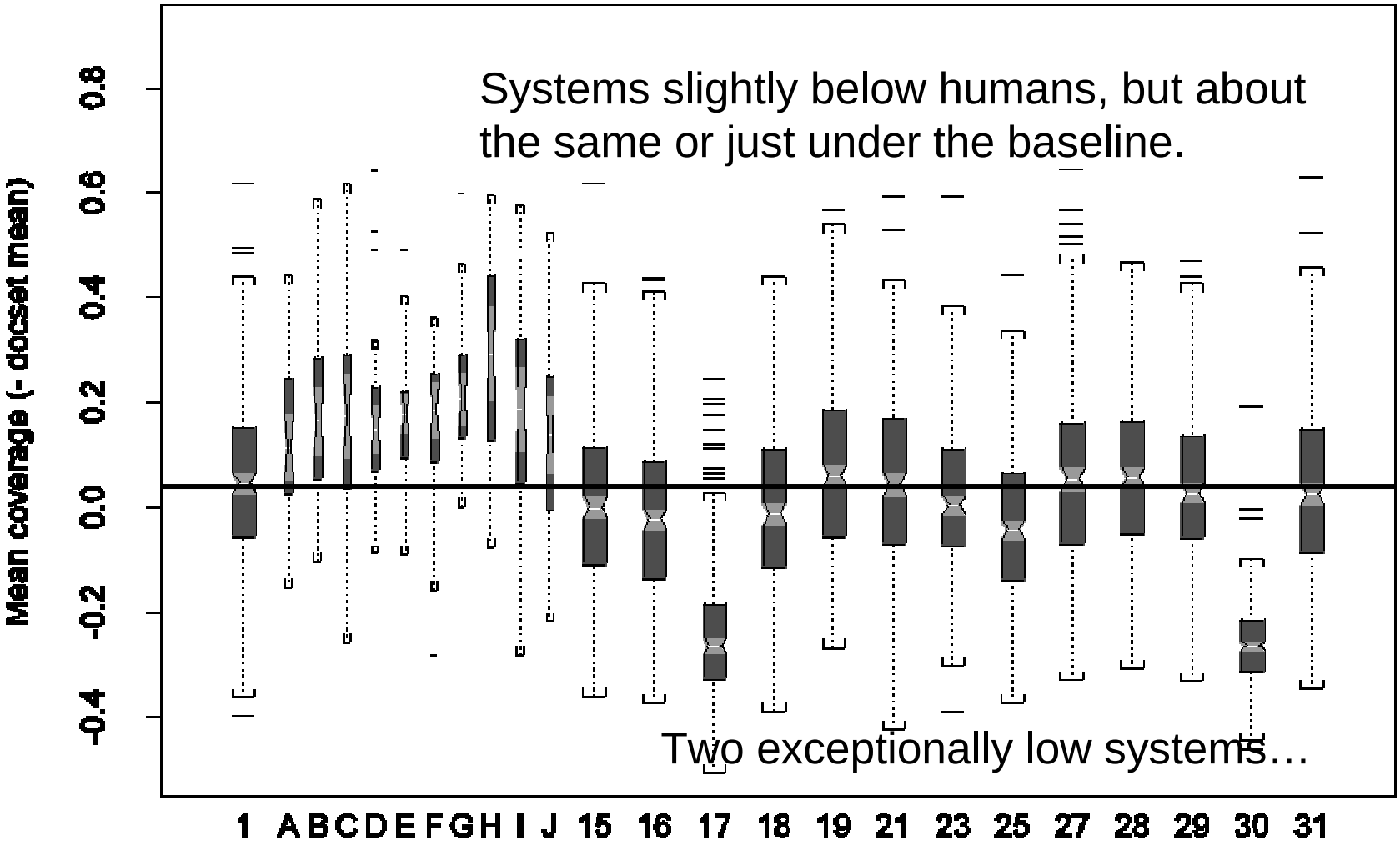


Mean coverage by peer source

Multi-document abstracts by size

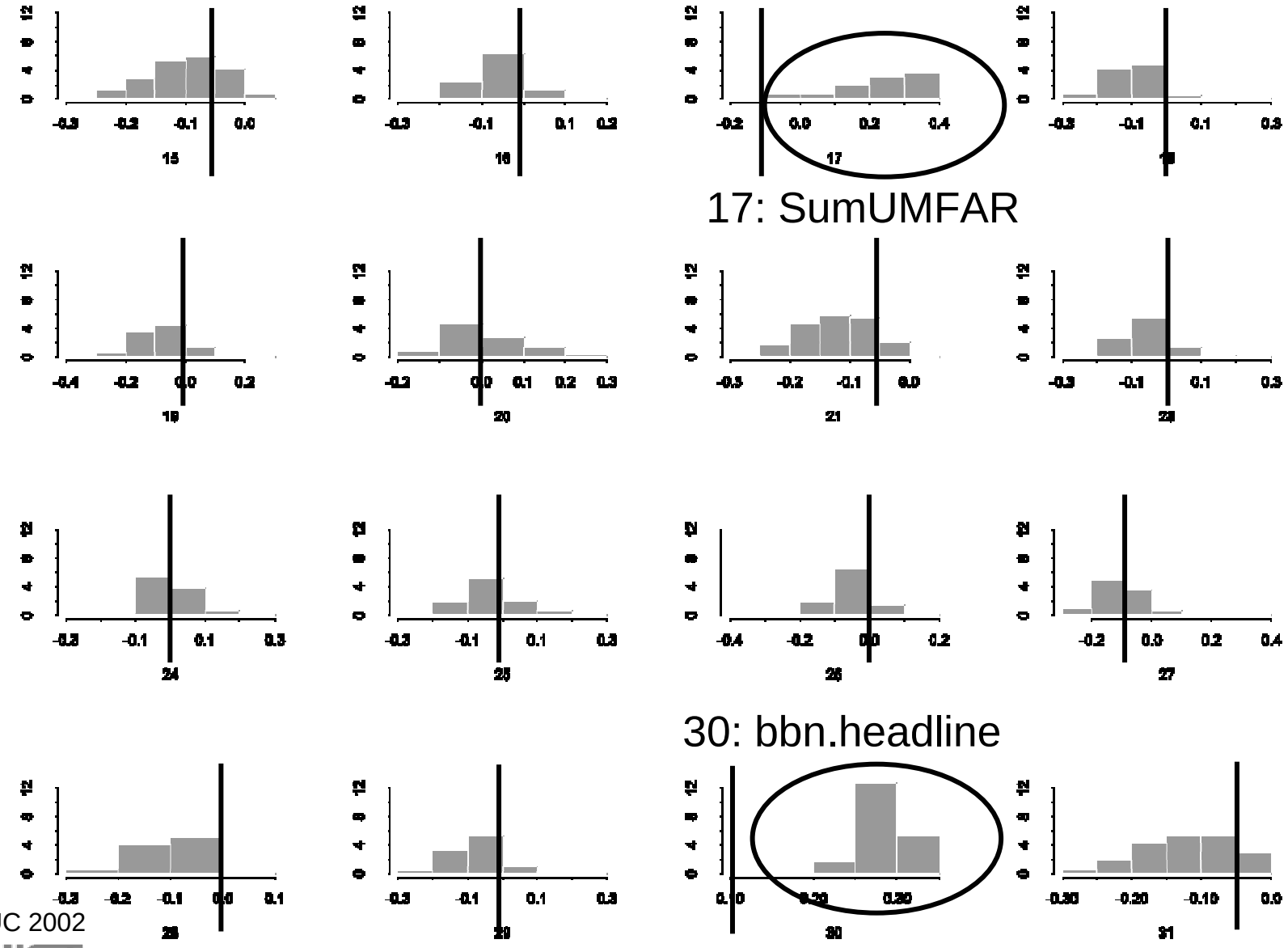


Comparing systems
(mean coverage - docset mean for all peers)
Single-document abstracts



What systems benefited from length-adjustment?

Distribution of (adjusted coverage – coverage)

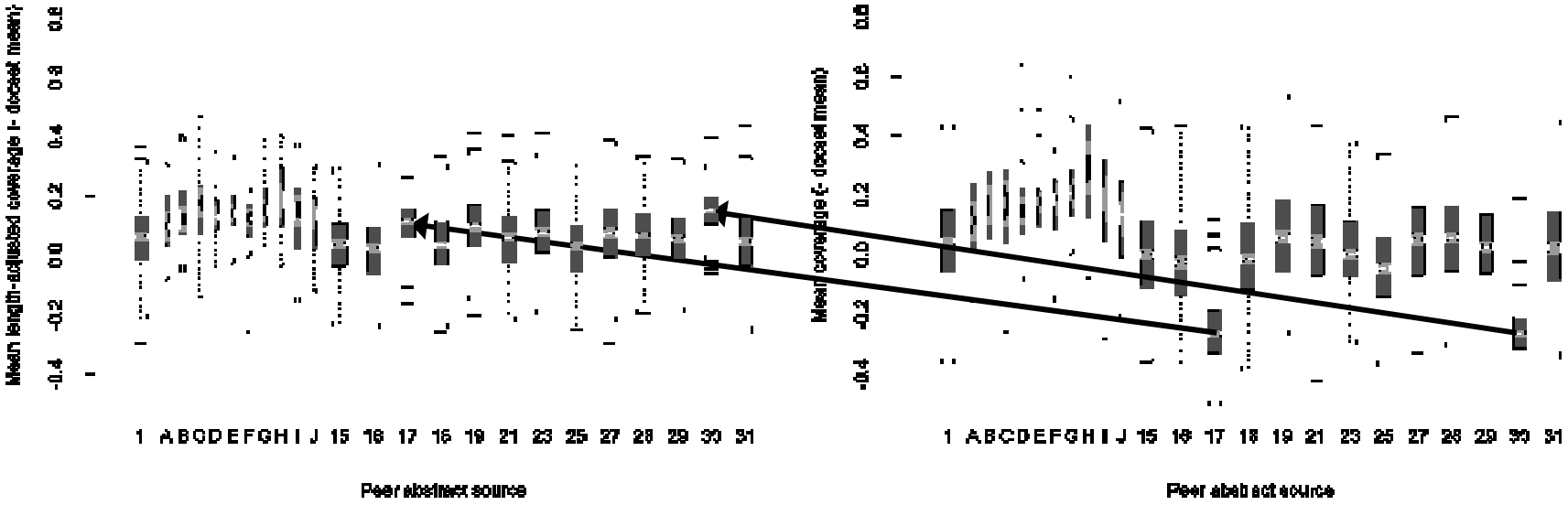


17: SumUMFAR

30: bbn.headline

Effect of length-adjustment by system
Single-document abstracts

- Using the a length adjustment seems to work: rewards shorter summaries with respect to longer
- Appropriate amount of boost is application-dependent

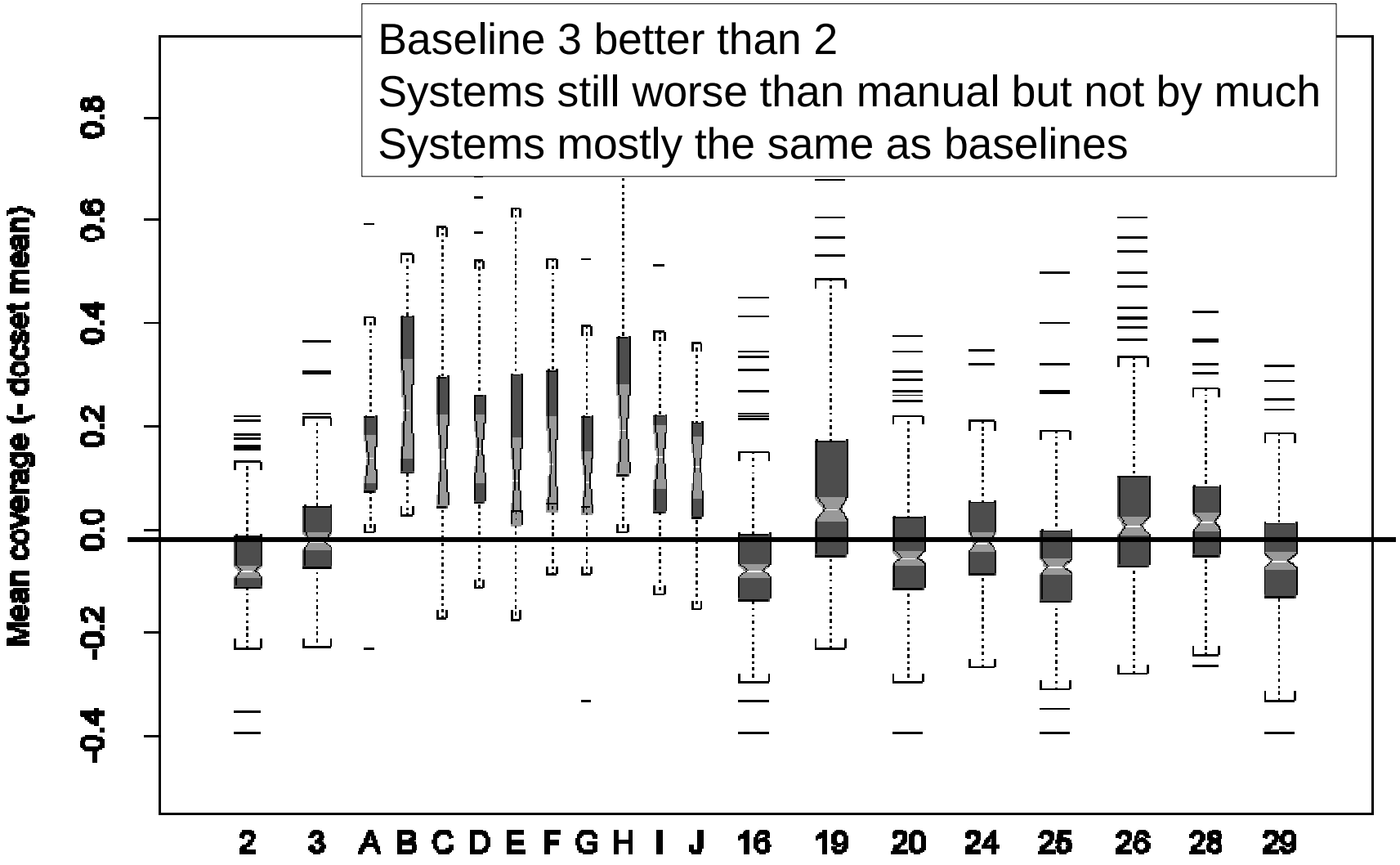


Length-adjusted

Not length-adjusted

Comparing systems (mean coverage - docset mean)

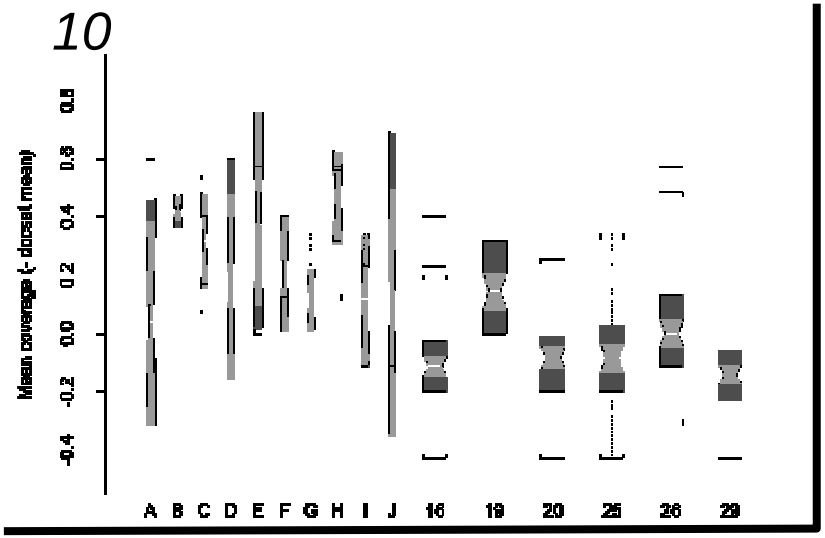
Multi-documents, all sizes



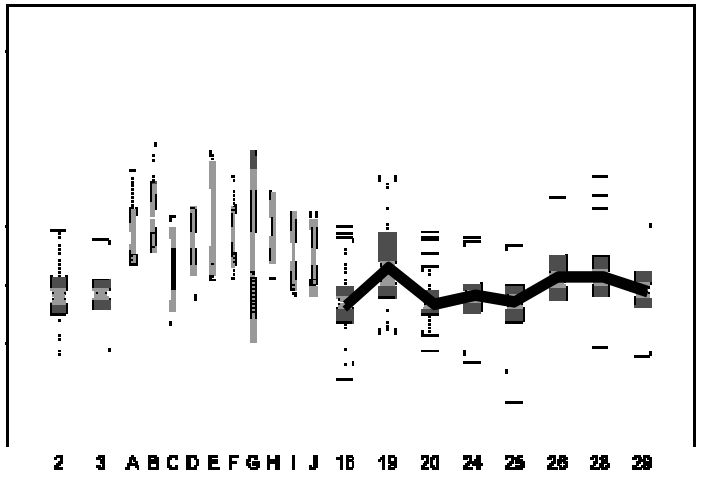
Comparing systems (mean coverage - docset mean)

Multi-document abstracts by size

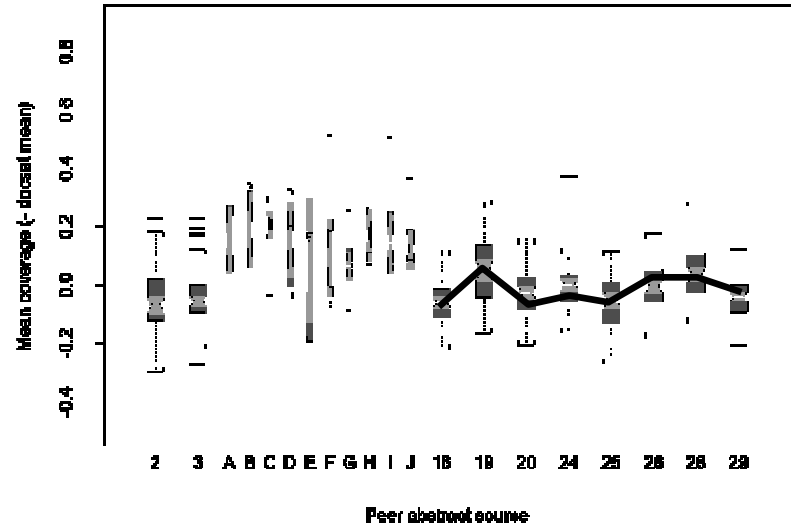
10



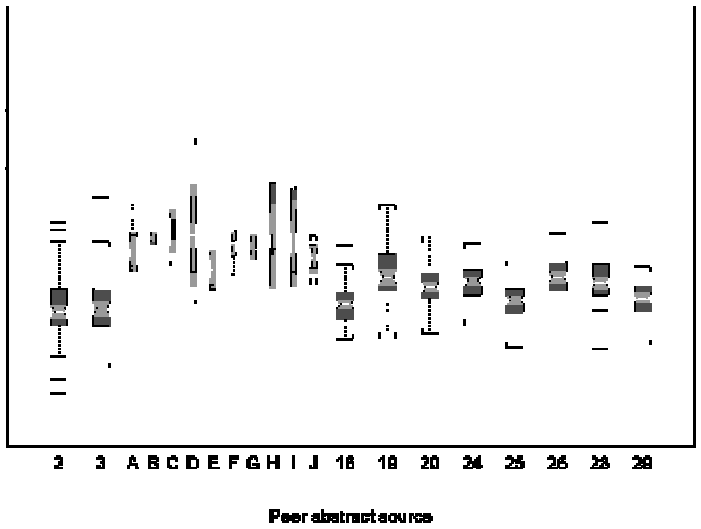
50



100



200



Gradual increase in most scores with abstract size

Relations among systems relatively stable

Factors affecting coverage
Analysis of Variance (ANOVA)

- Try ANOVA using simple model equation to see which factors matter: Mean coverage =
  - Grand mean +
  - System effect +
  - Document set (assessor) effect +
  - Noise
- Useful if
  - Other main effects are small
  - Interactions are small
  - ANOVA assumptions mostly met
- Note: for system rankings,
  - Main effects of document set, assessor, etc are balanced
  - Only interactions can cause problems

ANOVA results for multi-doc summaries by length

(Manual summaries combined)

200-word

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
system	10	2.262390	0.2262390	37.67023	0
doc.set	56	3.300333	0.0589345	9.81297	0
Residuals	560	3.363235	0.0060058		

100-word

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
system	10	2.275274	0.2275274	21.53187	0
doc.set	56	4.010818	0.0716217	6.77786	0
Residuals	560	5.917524	0.0105670		

50-word

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
system	10	2.960068	0.2960068	19.73176	0
doc.set	56	5.827342	0.1040597	6.93660	0
Residuals	560	8.400861	0.0150015		

10-word

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
system	6	8.12534	1.354223	32.92919	0.0000
doc.set	56	3.96627	0.070826	1.72220	0.0019
Residuals	336	13.81811	0.041125		

System and docset/assessor are significant

Account for most of the variability.

Other main factors and interactions…

- Main effect of document set type, document selector, summary author, etc. are not distinguishable from the noise
- Interactions are expected, but the experimental design lacks replicates – the main basis for directly estimated interactions
- Multi-judgment (Phase 2) data provide basis for assessing interactions
  - Designed to gauge effect of different assessors
  - Restricted to 6 document sets
  - 3 assessors,
    - none of which selected the documents or summarized them
    - used the same models to evaluate the same peers

Multiple judgment (Phase 2) study…

- Interactions are present as expected, but not large

```
coverage = grand mean + assessor + system + doc.set +
           assessor:system + assessor:doc.set + system:doc.set

                  Df Sum of Sq    Mean Sq  F Value      Pr(F)
          assessor  7 0.4790834 0.06844049 38.91456 0.00000000
            system 10 0.7434290 0.07434290 42.27061 0.00000000
           doc.set  5 0.1695724 0.03391449 19.28343 0.00000001
 assessor:system 70 0.3729502 0.00532786  3.02937 0.00061792
assessor:doc.set  3 0.0183594 0.00611979  3.47965 0.02797229
  system:doc.set 50 0.3669116 0.00733823  4.17244 0.00003924
       Residuals 30 0.0527621 0.00175874
```

…Multiple judgment (Phase 2) study

- Table of interaction sizes for system:docset can provide starting points for study:

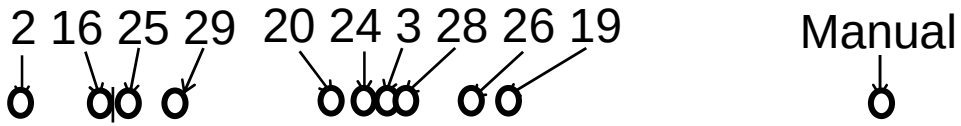
	D070	D071	D081	D094	D099	D102
16	0.0375	-0.1255	0.0005	0.0366	0.0476	0.0193
19	0.0845	0.0589	-0.1333	-0.1200	0.1243	-0.0173
2	-0.0174	-0.1077	-0.0579	0.0524	0.0253	0.0944
20	-0.0491	0.0646	0.0521	0.0233	-0.0718	-0.0256
24	0.0390	0.0004	0.0222	0.0418	-0.0757	-0.0455
25	-0.0694	0.0916	0.0313	-0.0180	-0.0781	0.0270
26	-0.0026	0.0007	-0.0066	-0.0119	0.0785	-0.0342
28	-0.0496	0.0677	0.0501	-0.0042	-0.0248	-0.0308
29	-0.0051	-0.0084	-0.0502	0.0440	0.0059	-0.0010
3	0.0378	-0.0368	0.1287	-0.0277	-0.0857	-0.0020
MANUAL	-0.0055	-0.0055	-0.0368	-0.0164	0.0543	0.0157

E.g., System 19 seems to have largish interactions with D081, 94, and 99. Why?

System differences using multiple comparisons..
(Tukey's)

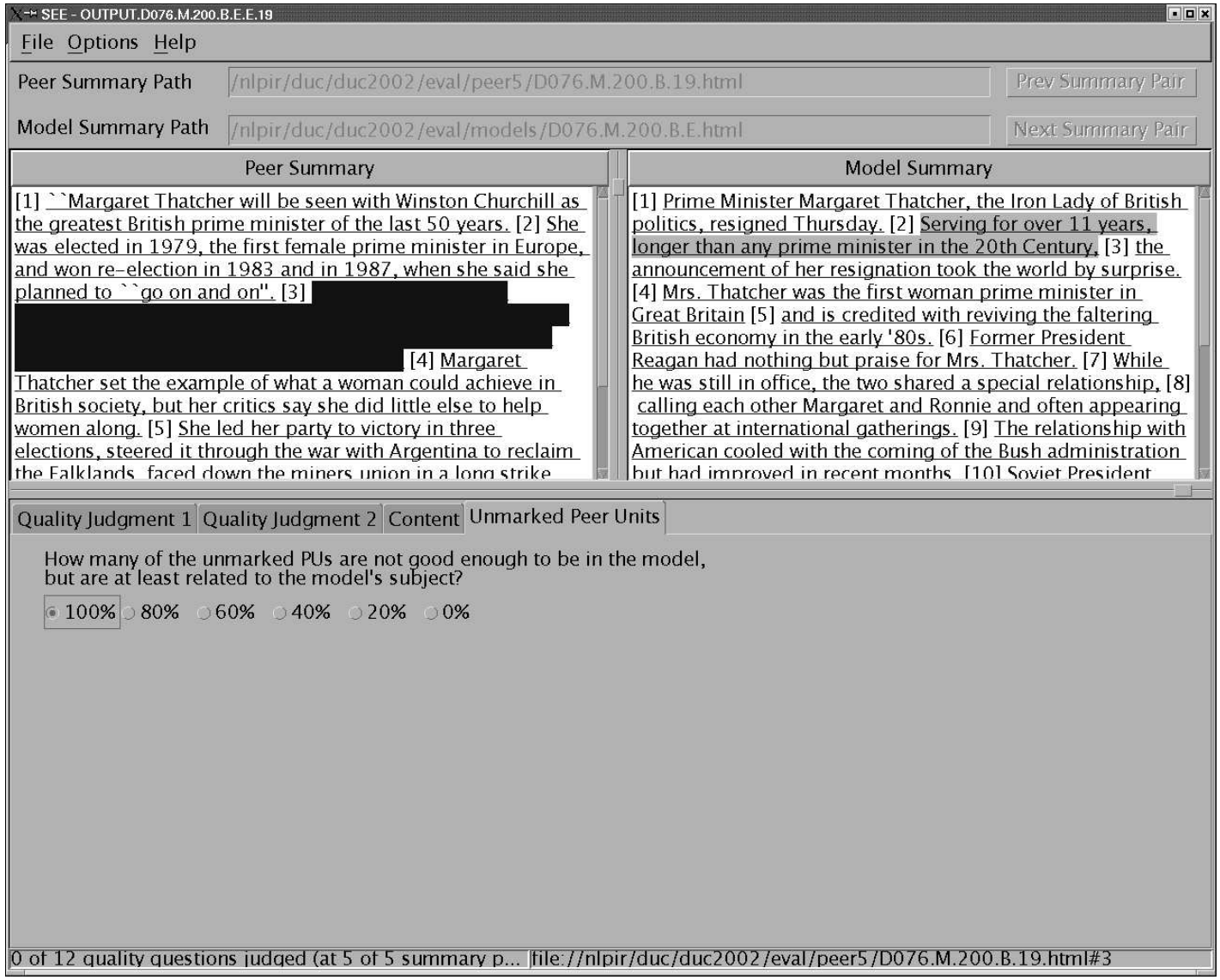
- Use multiple comparisons test to answers questions about real versus chance differences in baseline, system, and manual abstracts in terms of mean coverage

- For 200-word multi-doc abstracts: 3 main groups with some members of second group on the borderline with first



- As target size decreases,
  - noise increases
  - results blur
  - harder to tell if things are really different

SEE: unmarked peer units



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%

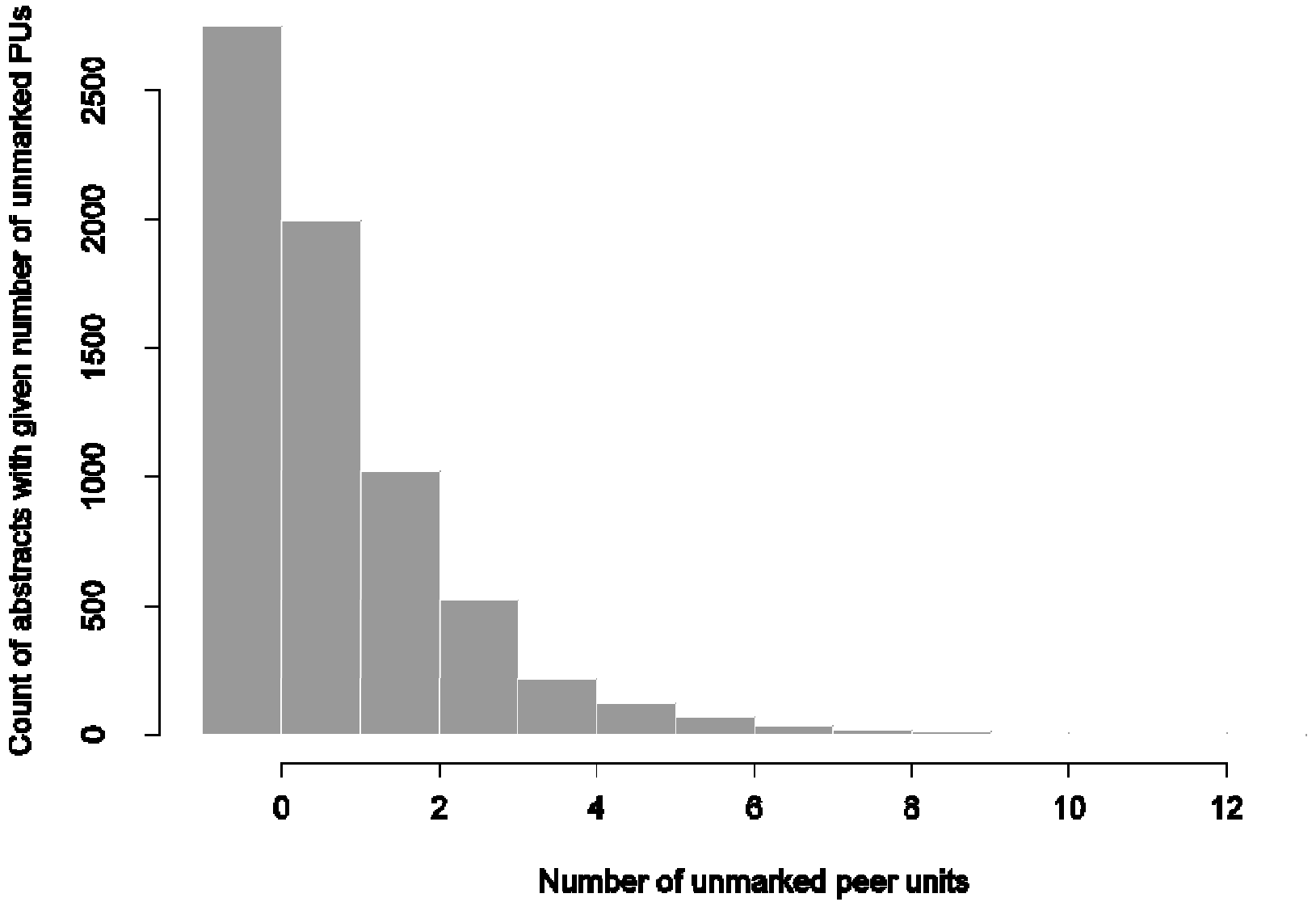
  4, choose 0, 20, 60, 80, or 100%

- If half the unmarked PUs are relevant

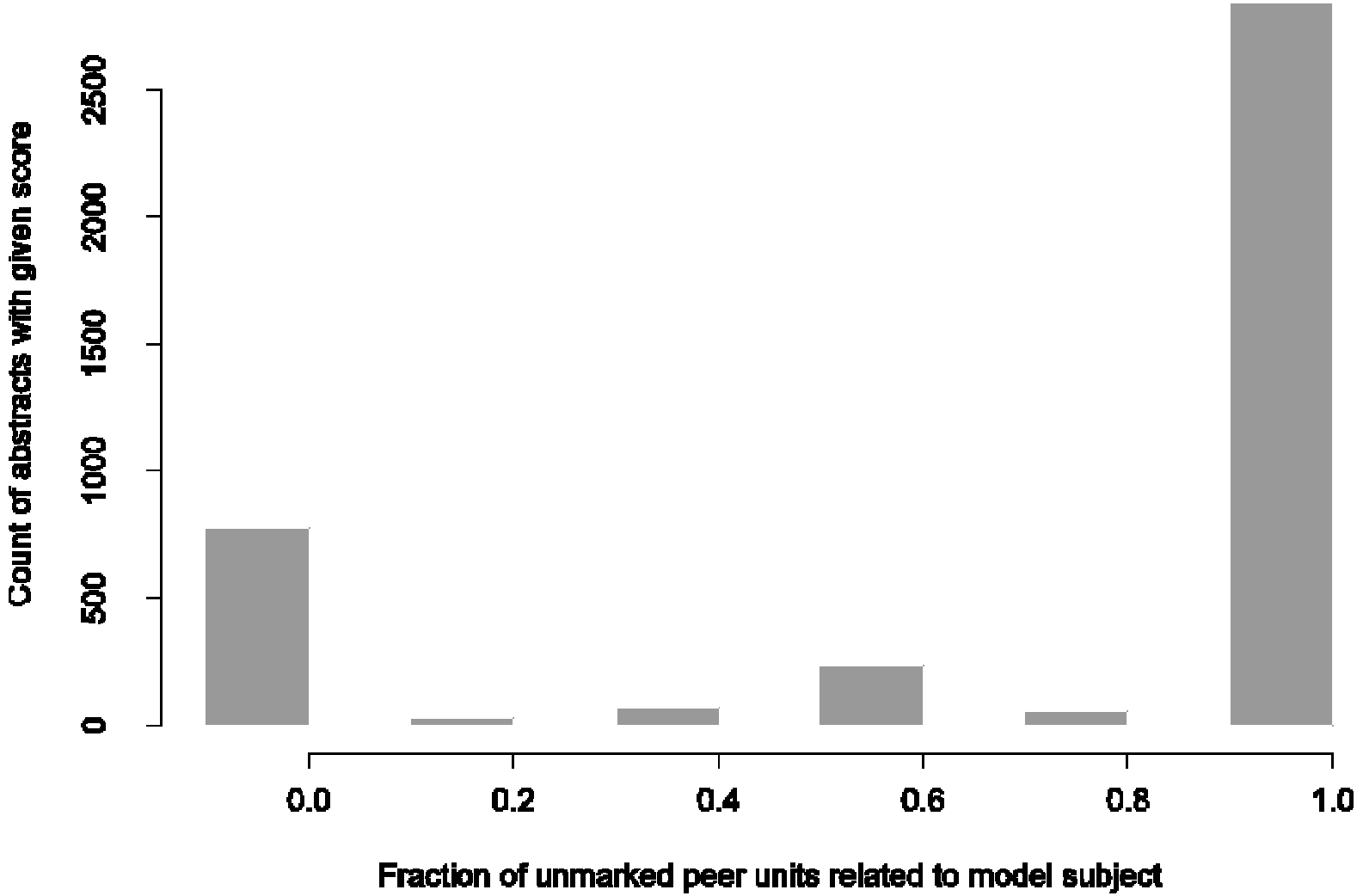
  Choose 60%

- Assessor feedback
  - Served to sanity check coverage
  - Some uncertainty about criteria for relatedness

How many unmarked peer units?



How many unmarked peer units related to model subject?



Summing up …

- Overall peer quality:
  - Appears assessors could handle the 12 peer quality questions
  - Results pass several sanity checks
  - Systems, baselines, and manual are distinguishable
  - But unintended “error” conditions were rare
- Per-unit content (coverage):
  - Surprising stability in system rankings across target sizes
  - Some systems stand out – why?
  - Room for improvement despite disagreement among humans
  - Too many systems are no better than baselines
  - Large number of MUs with no coverage needs further analysis
- Unmarked peers:
  - Cases of unmarked PUs being UNRELATED are rare
  - Should be examined

Incremental improvements…?

- Overall peer quality
  - Did any of the quality questions provide useful feedback?
  - If so, which ones?
  - Should others be substituted?
- Per-unit content (coverage)
  - Number of target lengths could be reduced
  - Still problems with EDU – sentence matching
    - Replace sentence separator with state-of-the-art
    - Better control EDU post-editing
  - All model units are not equal?
    - Investigate ways of categorizing/ranking MUs
    - Considering comparing MUs across target lengths to get simple 4-level ranking
- Unmarked peer units
  - Doesn't appear to be very informative