Overview of DUC 2005

Hoa Trang Dang

National Institute of Standards and Technology

Overview

- Direction of DUC
- Overview of DUC 2005 task
  - Task: documents, topics
  - Evaluation: measures, procedures, manual summaries
- Results of DUC 2005 evaluation
  - Relative performance of peers based on various measures
  - Relation between measures of content (Responsiveness, ROUGE)

Document Understanding Conferences (DUC)

Summarization roadmap created in 2000, progress from:

- simple genre $\rightarrow$ complex genre
- simple tasks $\rightarrow$ demanding tasks
  - extract $\rightarrow$ abstract
  - single document $\rightarrow$ multiple documents
  - English $\rightarrow$ other language
  - generic summaries $\rightarrow$ focused or evolving summaries
- intrinsic evaluation $\rightarrow$ extrinsic evaluation

DUC 2001-2004 investigated summarising

- for single documents, multi-documents
- for news material
- at various lengths
- of various sorts including generic author-reflecting, viewpoint-oriented, novelty capturing, query-oriented
- comparing system summaries with manual ones, and (automatic) baseline ones
- using a range of evaluation criteria and performance measures including:
  - intrinsic measures: quality, coverage of reference summary content units (SEE), ngram coincidence with reference summary (ROUGE)
  - extrinsic measures (simulated): usefulness and responsiveness.

Past intrinsic evaluation of content (SEE Coverage)

- Compare each peer against a single model summary
- Segment model summary into model units (MUs=EDU) (Soricut and Marcu, 2003)
- For each Model Unit:
  1. mark peer sentences that express any of the meaning in the MU
  2. the marked peer sentences together express [0, 20, 40, 60, 80, 100]% of the meaning expressed in the Model Unit.
- Mean coverage: average of the per-Model Unit judgments

Concerns:

- EDUs might be too big
- Variation in human summaries $\rightarrow$ Comparing against a single model summary gives unstable scores/rankings

Goals of DUC 2005

- **Develop new methods for evaluating summary content** that are
  - discriminative
  - stable (across assessors and/or sets of model summaries)
  - diagnostic
- Single task: multi-document, focused summaries requiring abstraction
- Control for variation in human summaries by specifying
  - Focus of summary
  - Simple user model/preference
- Create many human summaries, measure effect of variation on evaluation

DUC 2005 topics, document sets

50 sets of documents from *LA Times* and *Financial Times*

- 9 NIST assessors selected document sets and developed one topic statement per docset
- 25-50 documents per topic, average 32 documents per topic
- Documents in each docset contributed to answering the question(s) in the topic statement
- Topic indicated granularity of desired summary:
  - specific: describe specific instances of entities/events
  - general: generalize

Example topic statement

num: D345
title: American Tobacco Companies Overseas
narr: In the early 1990's, American tobacco companies tried to expand their business overseas. What did these companies do or try to do and where? How did their parent companies fare?
granularity: specific

Question-focused summarization task

- Given topic statement, granularity, document set
- Create fluent, 250-word answer to questions in topic statement, using information in document set
- Requires IR (passage level) + compression

Model summaries

- 10 assessors (including 9 topic developers): experienced information analysts, journalists, teachers
- 30 topics: 4 manual summaries each
- 20 topics: 9 manual summaries each
- About 5 hrs/summary

Example manual summary (specific)

As smoking lost acceptance in the United States, US cigarette producers fought to survive. They formed or joined multi-national conglomerates and pursued international tobacco markets as communism fell and monopolies were relaxed. Competition also was heavy in Western Europe, where anti-smoking policies and laws were just beginning.

Phillip Morris (with Marlboros) wooed Russian by supplying 20billion cigarettes after the 1990 tobacco riots and began planning and building factories and entering into joint ventures as the industry was privatized. It also moved into former Soviet states, building a factory in Lithuania and acquiring controlling of a Kazakhstan factory. Phillip Morris also brokered deals in Eastern Europe including Hungary (aiming to export to the Ukraine), Romania, Bulgaria, Czech Republic, Poland and eastern Germany. Asian markets, especially Japan, Taiwan and China, were also pursued. Phillip Morris' success was revealed by its increase in exports (24.4% in 1990), overseas sales gains (24% in 2d quarter 1991), and rising international tobacco profits (19.9% in 1992). In 1993, Phillip Morris had 12% of the 320 billion-cigarette market in Japan.

RJR(eynolds) Nabisco also diversified and invested overseas. It built factories in Turkey and Poland, bought factories in Germany and Hungary (also aiming at exports to the Ukraine), and entered joint ventures in Kazakhstan and China. Nabisco's net sales in Europe, Asia and the mid-East increased by 19% in 1991; international tobacco sales rose by $20million in 1992, and international tobacco sales were up by 12% in 1993. In 1991 it sold 14 billion cigarettes in Russia.

Example manual summary (general)

Single-sex education is superior for girls. They are able to study without intimidation and with fewer distractions. They may feel freer to take "traditional male" studies and do not feel forced to defer to the boys as they can feel in coeducational classes. The women are more likely to graduate and obtain advanced degrees. Both girls and boys can forge close friendships and camaraderie in single sex institutions. Boy's schools can pursue a sporty regimen that would too "macho" for girls. In the United Kingdom single-sex schools are the ones performing better academically with only one school in the top 30 being coeducational before the age of 16. Single-sex schools have few negatives effects for boys or girls. Some parents in California believe sex education classes should be segregated because the students would be more comfortable asking questions.

Single sex schools can encounter financial problems since they draw students from a smaller pool and providing dual systems can be expensive. Many parents dislike sex segregation and many students prefer coeducation. Girls may view an all-girl school to be a nunnery and few in the United States even consider enrolling in a women's college. Some believe coeducation better prepares children for life in the present generation. In the 1960's and 1970's single-sex schools were viewed as a barrier to successful socialization. Students in California preferred that the sex education classes be integrated and the dual classes are more expensive.

Participants and automatic runs in DUC 2005

Organization	RunID	Organization	RunID
(NIST Baseline)	1	The Hong Kong Polytechnic University	17
Chinese Academy of Sciences	2	Royal Institute of Technology KTH KOD	18
CL Research	3	Simon Fraser University	19
Columbia University	4	Thomson Legal & Regulatory	20
FreeText Software Technologies, Inc.	5	Toyohashi University of Technology	21
Fudan University	6	Universidad Autonoma de Madrid	22
IDA Center for Computing Sciences	7	University College Dublin	23
International Institute of Information Technology	8	University of Edinburgh	24
Institute for Infocomm Research	9	University of Karlsruhe	25
Information Sciences Institute (Daume)	10	University of Lethbridge	26
Information Sciences Institute (Lin)	11	University of Maryland and BBN	27
ITC-irst	12	University of Michigan	28
Laris/Larim Laboratory	13	Universite de Montreal	29
Language Computer Corporation	14	University of Ottawa	30
National University of Singapore	15	Technical University of Catalonia (UPC)	31
Oregon Health & Science University	16	University of Sheffield	32

Baseline: First 250 words of most recent document

Evaluation methods

- Linguistic quality (manual, intrinsic)
- Content coverage
  - Responsiveness (manual, pseudo-extrinsic)
  - ROUGE (automatic, intrinsic)
  - Pyramids (manual, intrinsic)
  - Basic Elements (automatic, intrinsic)

Manual assessment

- 10 Assessors
- Primary: all 50 topics
  - One assessor per topic: Linguistic quality, responsiveness
  - Assessor usually the same as topic developer
- Secondary: 20 topics
  - One assessor per topic: Responsiveness
  - Assessor usually NOT the same as topic developer
- Assessor always one of the summarizers for the topic

Linguistic quality

- 5 linguistic quality questions
- Each quality graded on a 5-point scale

    A. Very good

    B. Good

    C. Barely acceptable

    D. Poor

    E. Very poor

Linguistic quality questions

Q1. Grammaticality: The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.

Q2 Non-redundancy: There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., “Bill Clinton”) when a pronoun (“he”) would suffice.

Linguistic quality questions

Q3. Referential clarity: It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.

Q4. Focus: The summary should have a focus; sentences should only contain information that is related to the rest of the summary.

Q5. Structure and Coherence: The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

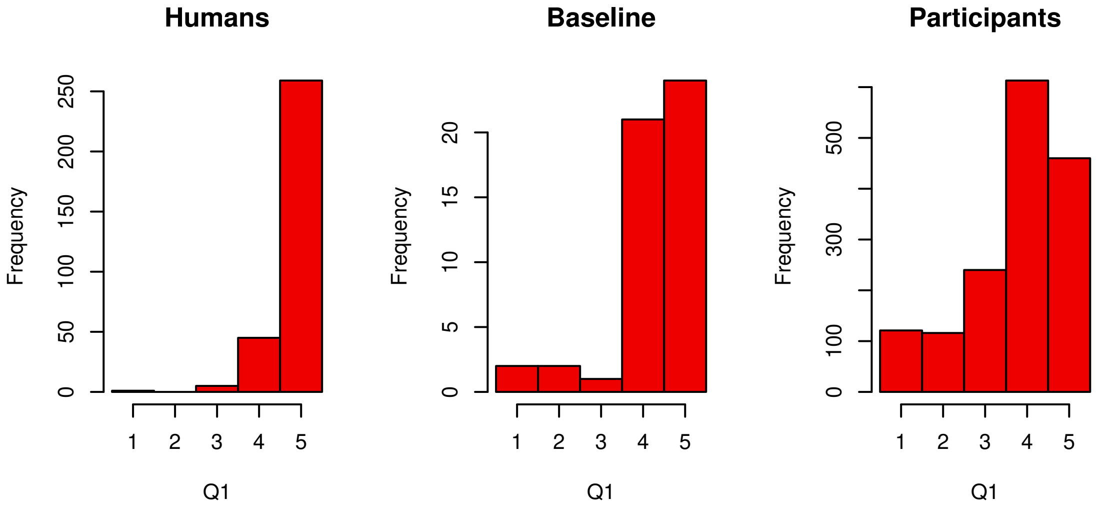
Responsiveness

- Responsiveness
  - Coarse ranking of all summaries (manual and automatic) per topic based on responsiveness of content to topic
  - Group all summaries into 5 groups: Worst=1, Best=5
  - At least one summary per group
  - Varying number of summaries per group
- Scaled responsiveness
  - Convert raw responsiveness into rank per topic
- Average scaled responsiveness = average rank across topics

ROUGE-1.5.5

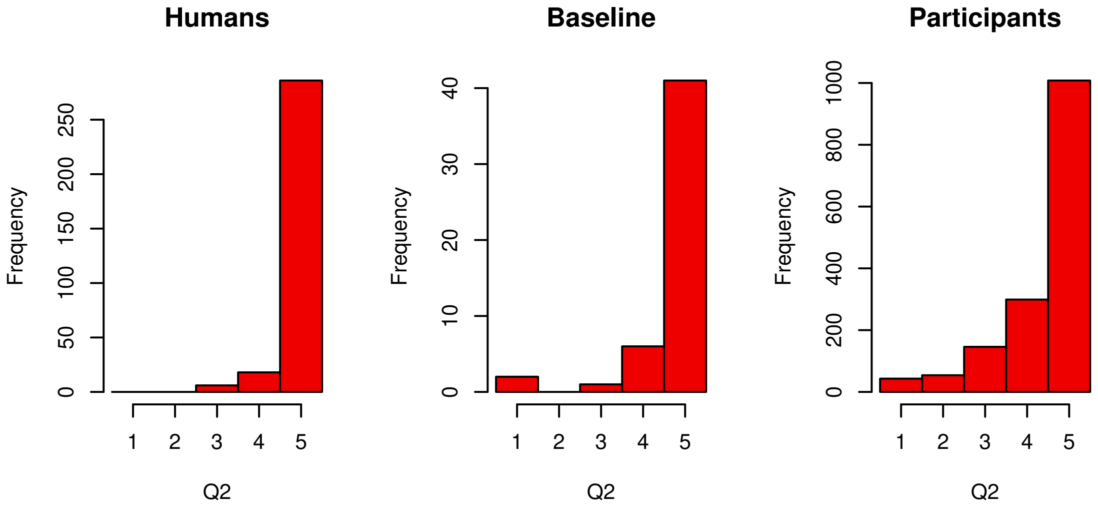
- Match n-grams between each peer summary and set of model summaries
- Weight n-gram by number of model summaries that it appears in
- Recall-oriented (precision, f-measure also available)
- Official DUC 2005 ROUGE metrics:
  - stem words before matching
  - ROUGE-2: match word bigrams
  - ROUGE-SU4: match skip bigrams, with skip distance of up to 4 words: "the dog" = "the quick scary little brown dog"
  - implement jackknifing for each (peer, topic) pair

Q1: Grammaticality



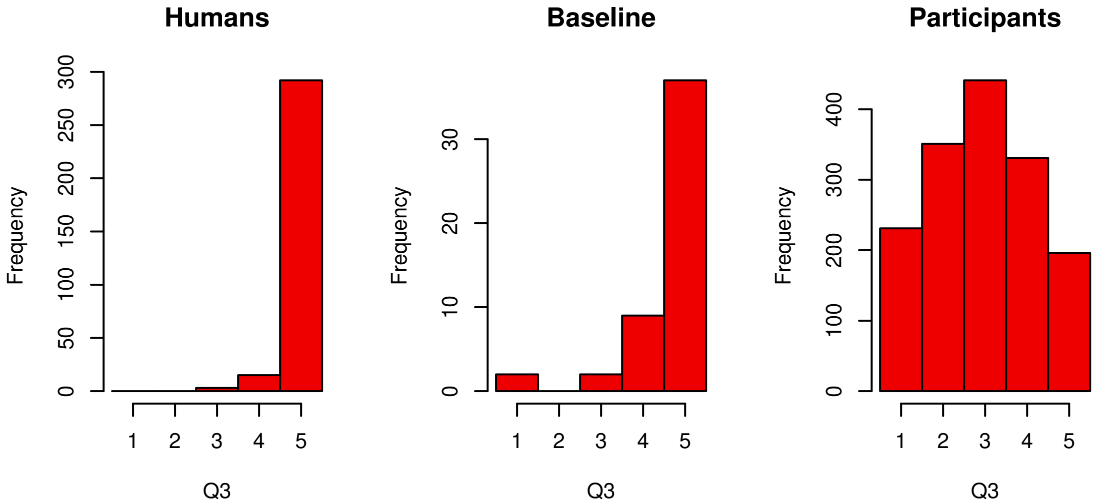
RunID	better than (Friedman's test)
5, 14, 16	10, 27, 15, 11, 31, 26, 25, 7
1, 20	10, 27, 15, 11, 31, 26, 25
18	10, 27, 15, 11, 31, 26
4, 28, 32	10, 27, 15, 11, 31
2, 6, 12, 22, 30	10, 27, 15, 11
3, 8, 9, 13, 17, 19, 21, 23, 24, 29	10, 27, 15
7	10

Q2: Non-redundancy



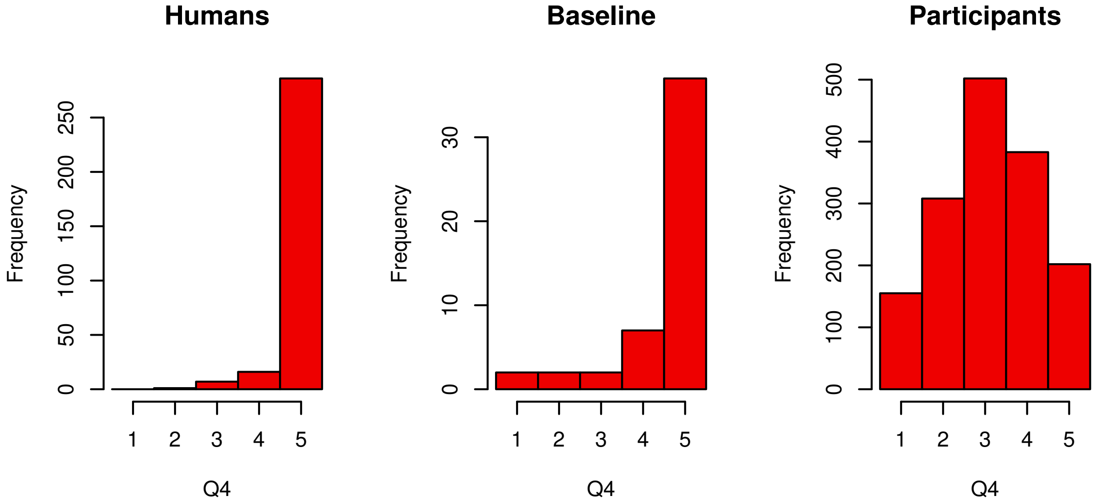
RunID	better than (Friedman's test)
1	15, 10, 31
21	15, 10
2, 7	15

Q3: Referential clarity



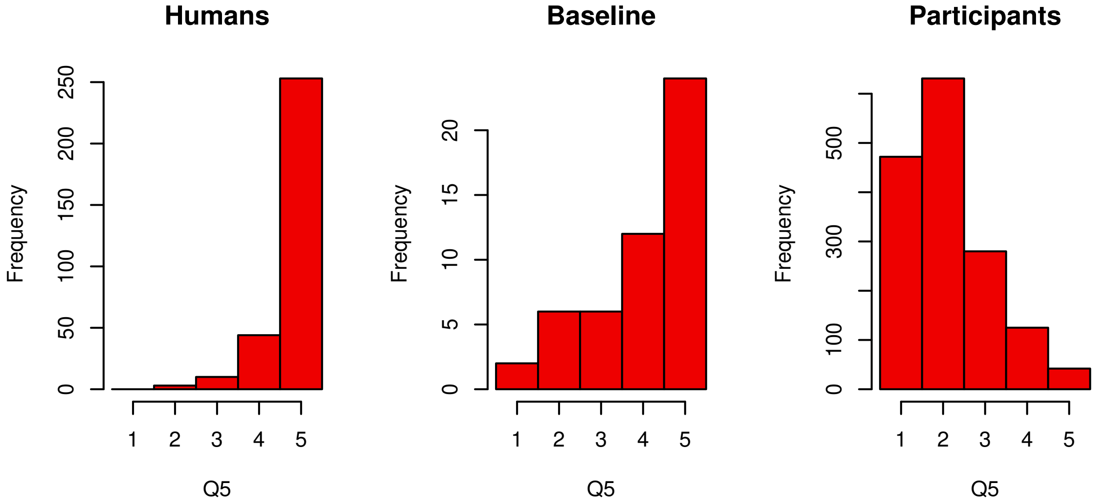
RunID	better than (Friedman's test)
1	all systems except 12
12	all systems except 1, 28, 17, 11
28	30, 6, 22, 24, 18, 13, 23, 31, 8
17	30, 6, 22, 24, 18, 13, 23, 31
11, 14, 21, 29	30, 6, 22, 24, 18, 13, 23
2, 5, 7, 32	30, 6, 22, 24
10	30, 6
4, 9, 15, 16, 26	30

Q4: Focus



RunID	better than (Friedman's test)
1	all systems except 12
12	23, 30, 22, 31, 27, 13, 18, 7, 11, 6, 9, 25, 10, 19, 21
2	23, 30, 22, 31, 27, 13, 18
4, 17	23, 30, 22, 31, 27, 13
14	23, 30, 22, 31
3, 5, 8, 15, 16, 29, 32	23, 30, 22
24, 26	23, 30
20, 28	23

Q5: Structure and coherence



RunID	better than (Friedman's test)
1	all systems except 12
12	all systems except 1, 2, 17, 14, 5, 28, 29
2	27, 13, 22, 30, 23, 31, 18, 10, 6, 11, 19
17	27, 13, 22, 30, 23, 31, 18
14	27, 13, 22, 30, 23, 31
5, 28, 29	27, 13, 22, 30, 23
16	27
4	27

Responsiveness

RunID	better than (Friedman's test)
(all humans	all automatics)
10	23, 31, 26, 1, 22, 30, 13, 2
5	23, 31, 26, 1, 22, 30, 13, 2
4	23, 31, 26, 1, 22, 30, 13
15	23, 31, 26, 1, 22, 30
29	23, 31, 26, 1, 22
11	23, 31, 26, 1, 22
17	23, 31, 26, 1, 22
8	23, 31, 26, 1, 22
7	23, 31, 26
14	23, 31, 26
6	23, 31, 26
28	23
21	23
19	23
24	23
9	23
16	23
32	23
12	23
25	23
18	23
27	23
20	23
3	23
2	23
13	23
30	23
22	23
1	23
26	23

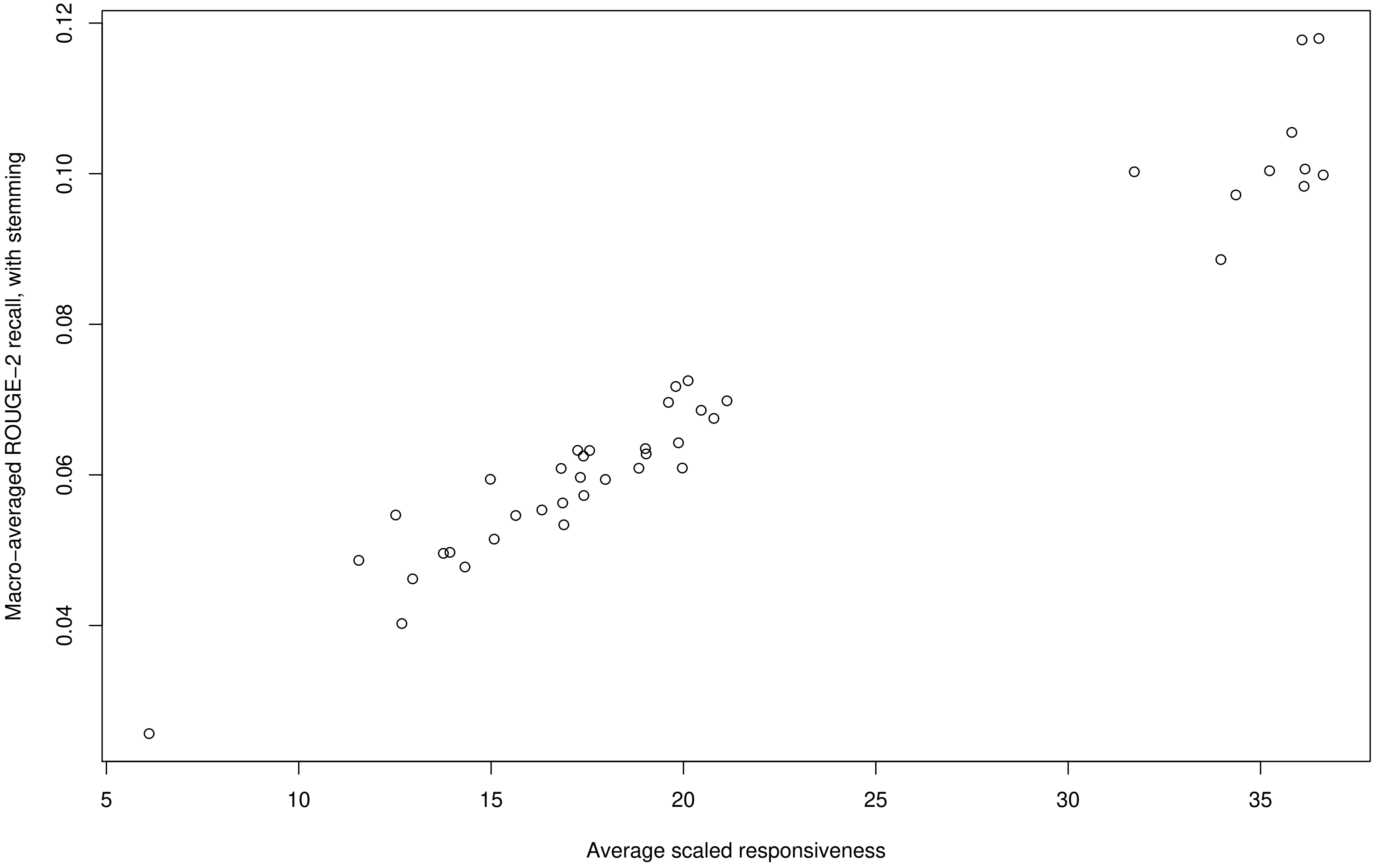
ROUGE-2 recall (ANOVA)

RunID	better than
C	all automatic, H, F, G
A	all automatic, H, F
I, B, J, E, D, G, F	all automatic
H	all automatic except 15
15	23, 1, 22, 2, 31, 30, 13, 20, 32, 27, 26, 18, 12
17	23, 1, 22, 2, 31, 30, 13, 20, 32, 27, 26, 18, 12
10	23, 1, 22, 2, 31, 30, 13, 20, 32
8	23, 1, 22, 2, 31, 30, 13, 20, 32
4	23, 1, 22, 2, 31, 30, 13, 20
5	23, 1, 22, 2, 31, 30, 13, 20
11	23, 1, 22, 2, 31
14	23, 1, 22, 2
16	23, 1, 22, 2
19	23, 1, 22, 2
7	23, 1, 22
9	23, 1, 22
29	23, 1
25	23, 1
6	23, 1
24	23, 1
28	23, 1
3	23, 1
21	23, 1
12	23, 1
18	23
26	23
27	23
32	23
20	23
13	23
30	23
31	23
2	23
22	23

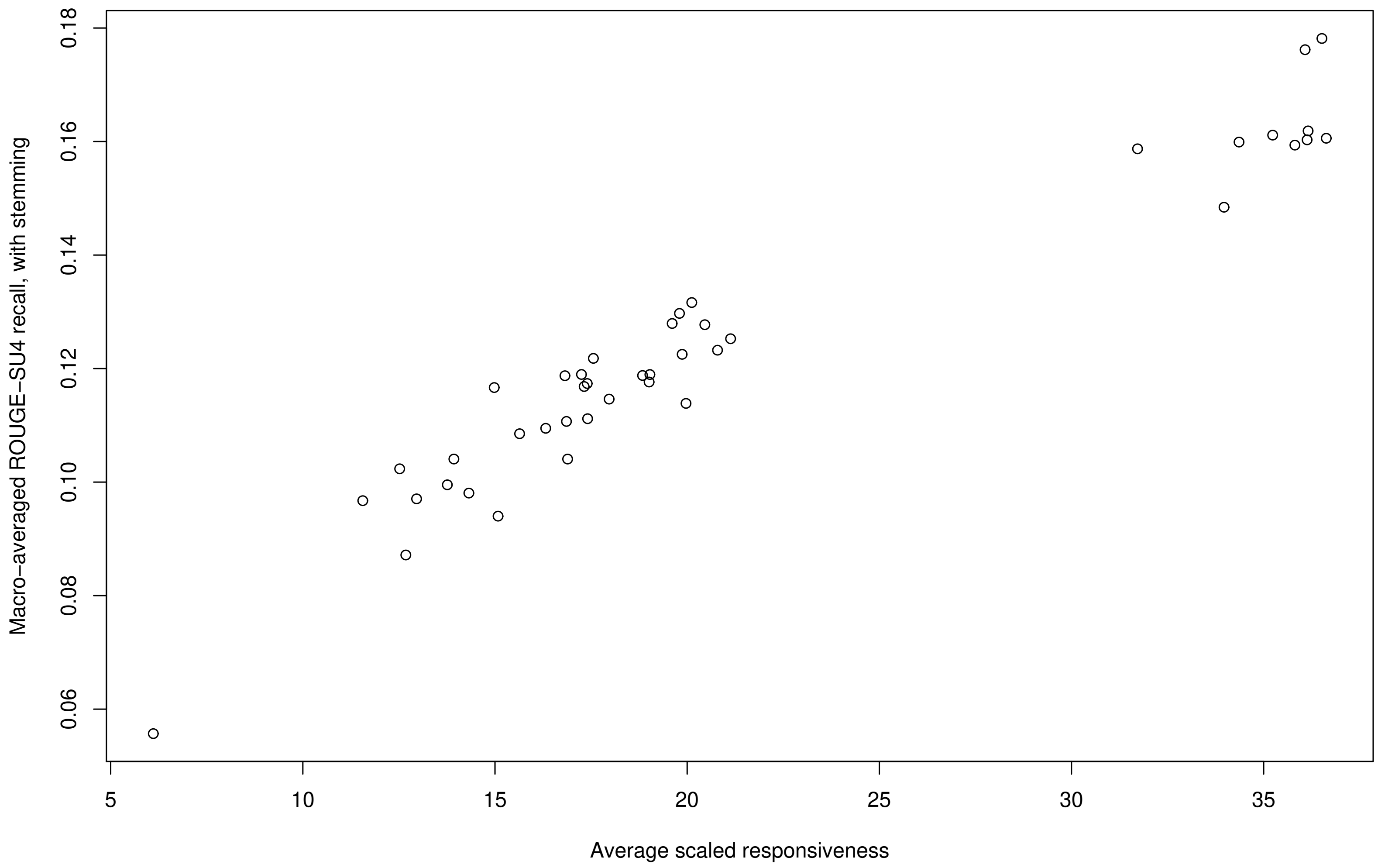
ROUGE-SU4 recall (ANOVA)

RunID	better than
C	all automatic, H, F, E
A	all automatic, H, F
I, J, B, G, D, E, F, H	all automatic
15	23, 1, 20, 31, 22, 2, 30, 26, 13, 32, 27, 18, 12, 21, 29, 28
17	23, 1, 20, 31, 22, 2, 30, 26, 13, 32, 27, 18, 12, 21
8	23, 1, 20, 31, 22, 2, 30, 26, 13, 32, 27, 18, 12, 21
4	23, 1, 20, 31, 22, 2, 30, 26, 13, 32, 27, 18, 12
10	23, 1, 20, 31, 22, 2, 30, 26, 13, 32, 27
5	23, 1, 20, 31, 22, 2, 30, 26, 13, 32
11	23, 1, 20, 31, 22, 2, 30, 26, 13, 32
19	23, 1, 20, 31, 22, 2, 30, 26, 13, 32
16	23, 1, 20, 31, 22, 2, 30, 26
7	23, 1, 20, 31, 22, 2, 30, 26
6	23, 1, 20, 31, 22, 2, 30
25	23, 1, 20, 31, 22, 2, 30
14	23, 1, 20, 31, 22, 2, 30
9	23, 1, 20, 31, 22, 2, 30
24	23, 1, 20, 31, 22, 2, 30
3	23, 1, 20, 31, 22, 2, 30
28	23, 1, 20, 31, 22
29	23, 1, 20, 31, 22
21	23, 1, 20
12	23, 1, 20
18	23, 1
27	23, 1
32	23, 1
13	23, 1
26	23
30	23
2	23
22	23
31	23
20	23
1	23

Average ROUGE-2 recall vs. Average Scaled Responsiveness



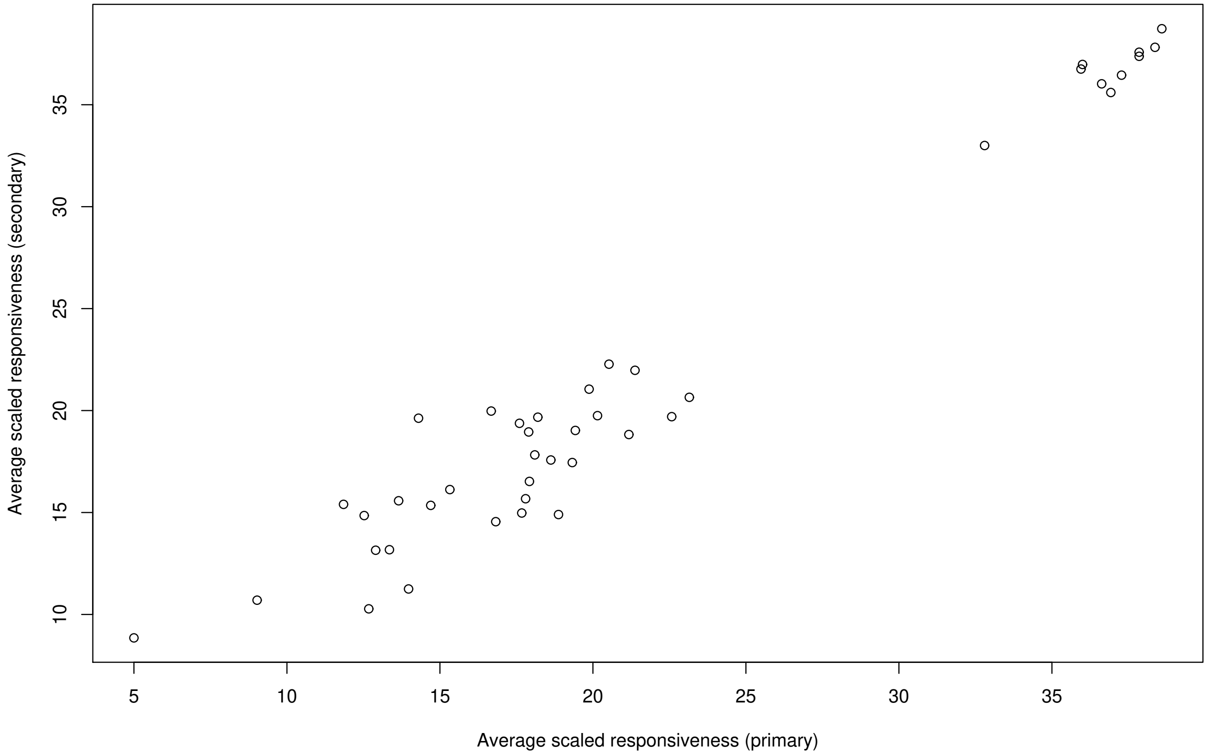
Average ROUGE-SU4 recall vs. Average Scaled Responsiveness



Correlation with average scaled responsiveness

Metric	Spearman	Pearson
Rouge-2 (all peers)	0.951	0.972 [0.953, 1.000]
Rouge-SU4 (all peers)	0.942	0.958 [0.930, 1.000]
Rouge-2 (automatic peers)	0.901	0.928 [0.872, 1.000]
Rouge-SU4 (automatic peers)	0.872	0.919 [0.855, 1.000]

Primary vs. Secondary Average Scaled Responsiveness



	Spearman	Pearson
All peers	0.900	0.976 [0.960, 1.000]
Automatic peers	0.775	0.822 [0.695, 1.000]

Conclusion

- Responsiveness
  - Humans indistinguishable
  - Humans $>$ Systems; 8 Systems $>$ Baseline $>$ 1 System
  - Stable, not diagnostic
- Rouge-2
  - Humans distinguishable
  - Humans $>$ Systems; 20 Systems $>$ Baseline $>$ 1 System
  - Correlates with responsiveness; Diagnostic?
- Rouge-SU4
  - Humans distinguishable
  - Humans $>$ Systems; 24 Systems $>$ Baseline $>$ 1 System
  - Correlates with responsiveness; Diagnostic?