Overview of the TAC 2008 Update Summarization Task

Hoa Trang Dang, Karolina Owczarzak

National Institute of Standards and Technology
U.S. Department of Commerce

Update Summarization Task

- Task
  - **main:** produce a 100-word summary from a set of 10 documents (Summary A)
  - **update:** produce a 100-word summary from a set of subsequent 10 documents, with the assumption that the information in the first set is already known to the reader (Summary B)

Update Summarization Task

- 48 topics
- 20 documents per topic in chronological order:
  - main summary (first 10 documents)
  - update summary (second 10 documents)
- 100 words per summary
- 4 model summaries
  - one summary by topic creator

Data

- AQUAINT-2 Corpus
  - part of LDC English Gigaword corpus 3rd Ed.
  - 2.5GB of text
  - news articles Oct 2004 – Mar 2006:
    - Agence France Presse
    - Xinhua News Agency
    - Los Angeles Times – Washington Post News Service
    - New York Times
    - Associated Press
- Average length of selected doc: 3368 wrds

Topics

- D0820D

Title: Submarine Rescue

Narrative: Describe efforts of the Russian navy to rescue the trapped submariners and any assistance provided by other countries. Include information regarding the results of the rescue mission and the results and consequences of the subsequent investigation into the matter.

Participants

- 33 teams
- 71 runs (up to 3 per team)
  - manual evaluation for 1st and 2nd priority runs (57)
  - automatic evaluation for all runs
- NIST baseline
  - first sentence(s) of the most recent document
  - up to 100 words

Manual Evaluation

- Overall Responsiveness

How well is the summary responding to the information need contained in the topic statement? How good is the structure of the summary and its linguistic quality?

- Overall Readability

What is the overall linguistic quality of the summary, independent of content? Note the fluency, structure, grammaticality, non-redundancy, referential clarity, focus, coherence.

Manual Evaluation

- Overall Responsiveness

1.....2.....3.....4.....5

Very Poor Poor Barely Acceptable Good Very Good

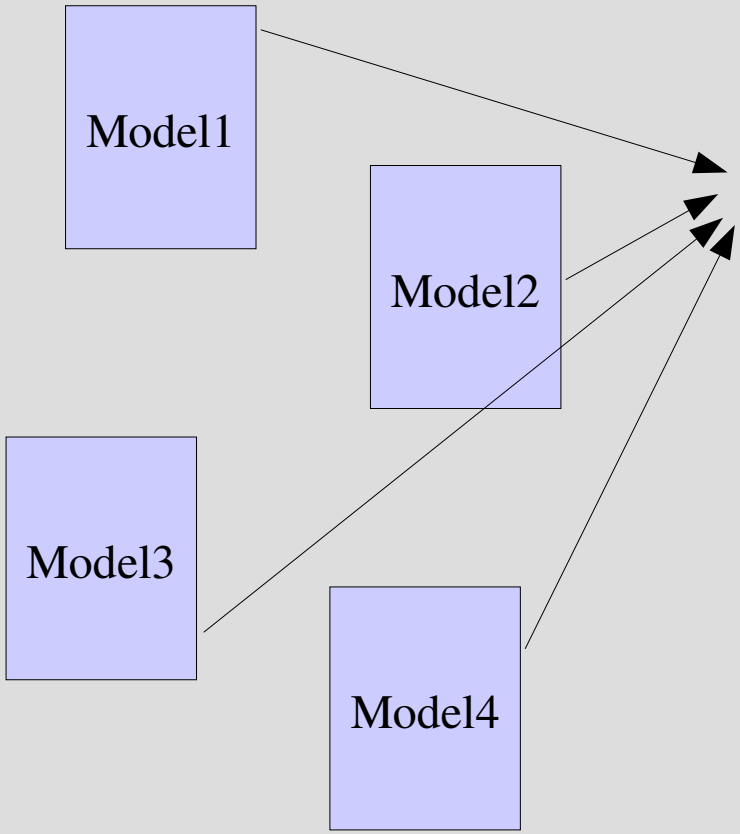
- Overall Readability

1.....2.....3.....4.....5

Very Poor Poor Barely Acceptable Good Very Good

Manual Evaluation

- Pyramid framework (Passonneau et al., 2005)



Summary Content Units (SCUs):

- Mini-submarine trapped underwater (4)
- Mini-sub snagged by underwater cables (3)
- Britain sent a robotic vehicle (3)
- U.S. sent underwater vehicles (2)
- Japan sent four vessels (2)
- British arrived first (2)
- Crew taken for medical examination (1)
- Military submarine (1)
- Mini-sub trapped in eastern Russia (1)
- U.S. sent equipment (1)

Manual Evaluation

- Pyramid framework (Passonneau et al., 2005)

SCU (4): Mini-submarine trapped underwater

contributor1: mini-submarine... became trapped... on the sea floor
contributor2: a small... submarine... snagged... at a depth of 625 feet
contributor3: mini-submarine was trapped... below the surface
contributor4: A small... submarine... was trapped on the seabed

Manual Evaluation

- Pyramid framework (Passonneau et al., 2005)

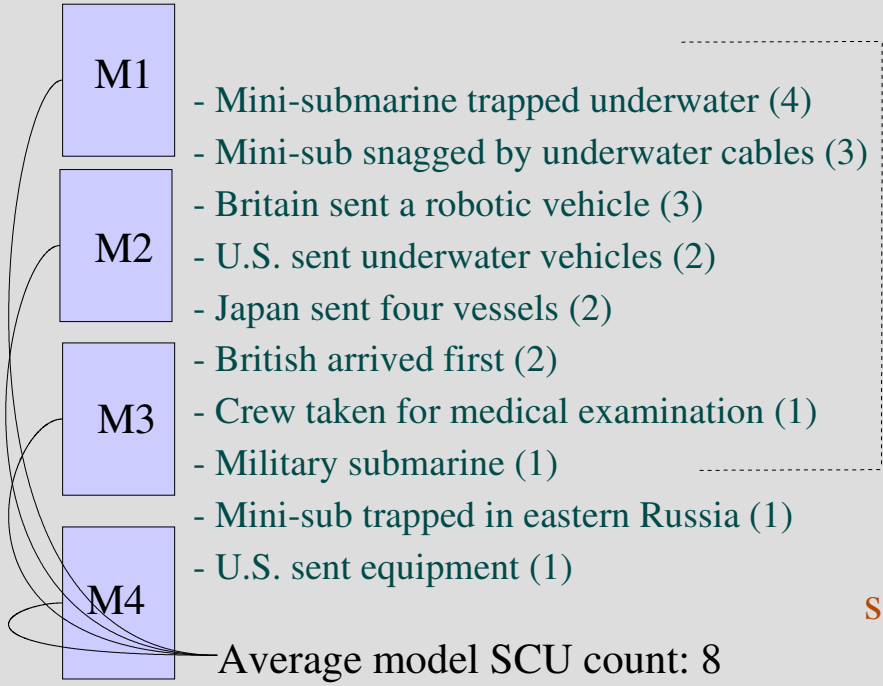
$$\text{score} = \frac{\text{total SCU weight}}{\text{max SCU weight possible with average SCU count}}$$

Candidate Summary

- Mini-submarine trapped underwater (4)
- Mini-sub trapped in eastern Russia (1)
- U.S. sent equipment (1)

Total SCU count: 3

Total SCU weight: 6



Max weight with 8 SCUs: 18

$$\text{score} = \frac{6}{18} = 0.33$$

Automatic Evaluation

- ROUGE (Lin, 2004)
  - ROUGE-2 recall: matching bigrams
  - ROUGE-SU4 recall: matching skip-bigrams (skip up to 4 intervening words)
- BE (Hovy et al., 2005)
  - BE-HM: matching head-modifier pairs

sent | call (obj)
sent | they (subj)
call | help (for)
help | international (mod)
sent | out (guest)

- Jackknifing for all metrics
  - evaluate each model summary against remaning 3 models
  - evaluate each automatic summary 4 times, each time against a different set of 3 models, average out

Results – Main vs Update

Macro-average per-topic scores

	Responsiveness		Readability		Pyramid	
	models	systems	models	systems	models	systems
Summaries A	4.620	2.324*	4.786	2.347	0.663	0.260*
Summaries B	4.625	2.024*	4.800	2.337	0.630	0.204*

	ROUGE-2		ROUGE-SU4		BE-HM	
	models	systems	models	systems	models	systems
Summaries A	0.117	0.079*	0.154	0.116*	0.078	0.038
Summaries B	0.117	0.068*	0.150	0.107*	0.089	0.039

* difference statistically significant with $p < 0.05$

Results – Models vs Systems

RESPONSIVENESS

D	4.833
F	4.729
G	4.708
A	4.688
B	4.583
H	4.583
C	4.500
E	4.354
23	2.667
49	2.667
44	2.635
50	2.625
14	2.615
11	2.542
24	2.521
52	2.479
25	2.479
41	2.479
37	2.479
26	2.469
6	2.469
51	2.448
1	2.427
13	2.427
42	2.417
45	2.385
34	2.385
2	2.385
12	2.344
46	2.333
17	2.323
19	2.312
43	2.260
3	2.240
35	2.219
10	2.219
15	2.208
22	2.198
54	2.188
48	2.177
4	2.167
36	2.156
16	2.115
5	2.104
33	2.104
29	2.083
0	2.073
55	2.073
57	2.073
20	2.062
27	2.052
32	2.031
21	2.021
40	1.990
56	1.948
31	1.938
53	1.917
30	1.917
28	1.740
7	1.688
47	1.656
8	1.542
38	1.510
18	1.479
39	1.417
9	1.198

READABILITY

D	4.917
F	4.896
G	4.854
A	4.833
B	4.812
E	4.729
H	4.688
C	4.604
0	3.333
49	3.073
23	2.958
50	2.896
52	2.896
24	2.885
26	2.885
51	2.812
44	2.792
25	2.771
34	2.760
1	2.719
14	2.708
46	2.646
6	2.594
17	2.562
37	2.552
45	2.521
13	2.479
16	2.458
10	2.448
31	2.438
33	2.438
35	2.427
5	2.427
4	2.417
22	2.406
11	2.406
27	2.375
15	2.365
20	2.354
2	2.354
47	2.344
3	2.333
41	2.323
53	2.302
54	2.292
57	2.281
36	2.240
48	2.208
19	2.188
21	2.177
56	2.156
12	2.031
42	2.031
32	2.010
43	2.000
40	1.958
30	1.938
55	1.833
29	1.802
39	1.771
18	1.760
7	1.677
9	1.635
28	1.625
38	1.448
8	1.312

PYRAMID

G	0.805
D	0.708
H	0.655
C	0.651
B	0.625
F	0.613
A	0.608
E	0.511
11	0.331
44	0.319
14	0.317
41	0.313
23	0.304
37	0.301
49	0.299
6	0.296
13	0.295
25	0.290
50	0.287
43	0.285
45	0.284
12	0.282
42	0.280
51	0.278
2	0.276
19	0.276
24	0.275
52	0.272
48	0.263
15	0.263
1	0.261
34	0.260
26	0.258
35	0.250
17	0.249
3	0.242
10	0.238
36	0.234
46	0.234
29	0.234
22	0.232
54	0.230
4	0.229
55	0.222
16	0.222
20	0.219
40	0.212
21	0.212
27	0.212
32	0.206
30	0.204
57	0.202
28	0.191
5	0.190
33	0.186
53	0.184
56	0.180
0	0.163
31	0.160
8	0.153
38	0.140
7	0.138
47	0.130
18	0.085
39	0.073
9	0.055

Results – Models vs Systems

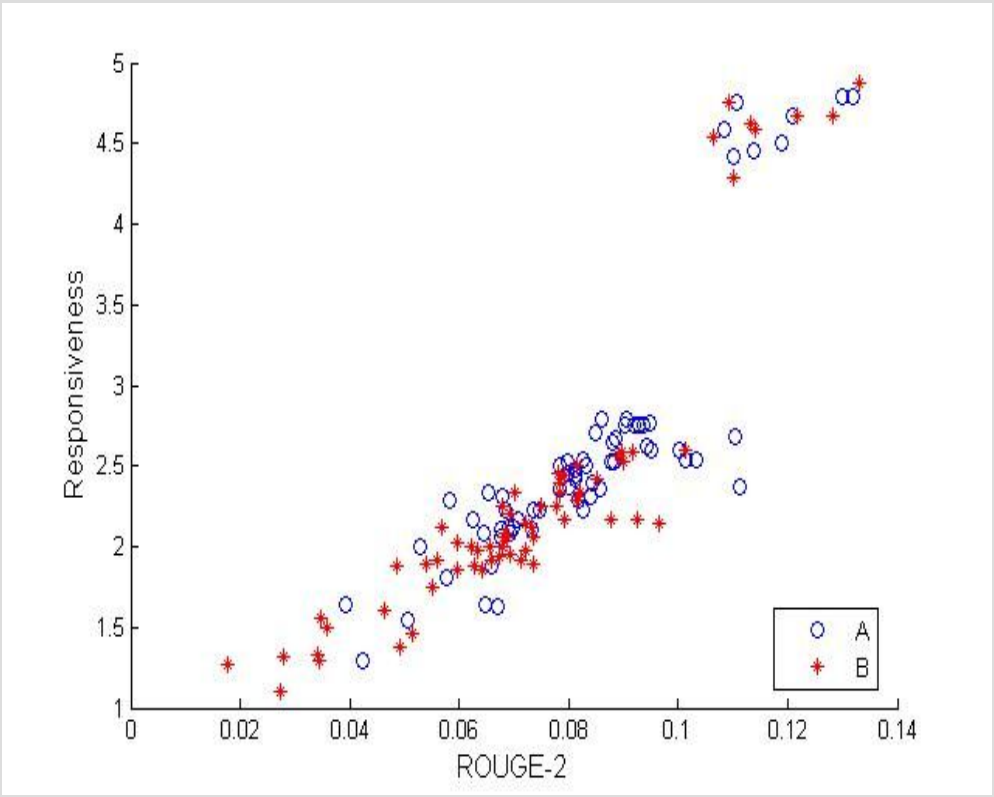
Macro-average submission scores

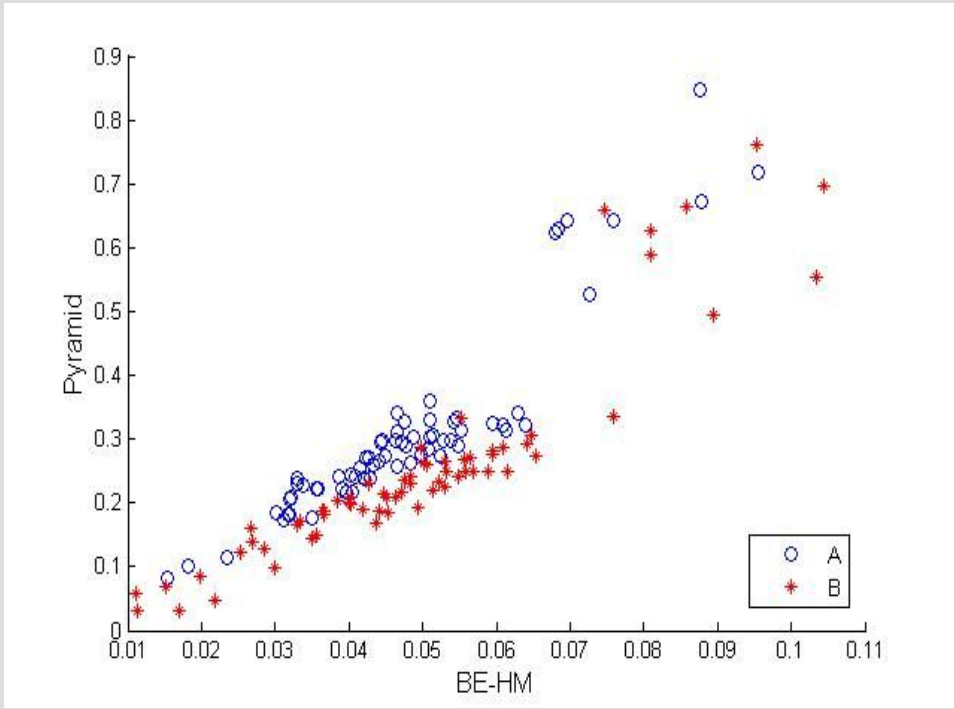
	Responsiveness	Readability	Pyramid
models	4.622*	4.792*	0.647*
systems	2.174*	2.342*	0.232*

	ROUGE-2	ROUGE-SU4	BE-HM
models	0.117*	0.152*	0.084*
systems	0.074*	0.111*	0.045*

* difference statistically significant with $p < 0.05$

Results – Models vs Systems





Manual Metrics - Correlation

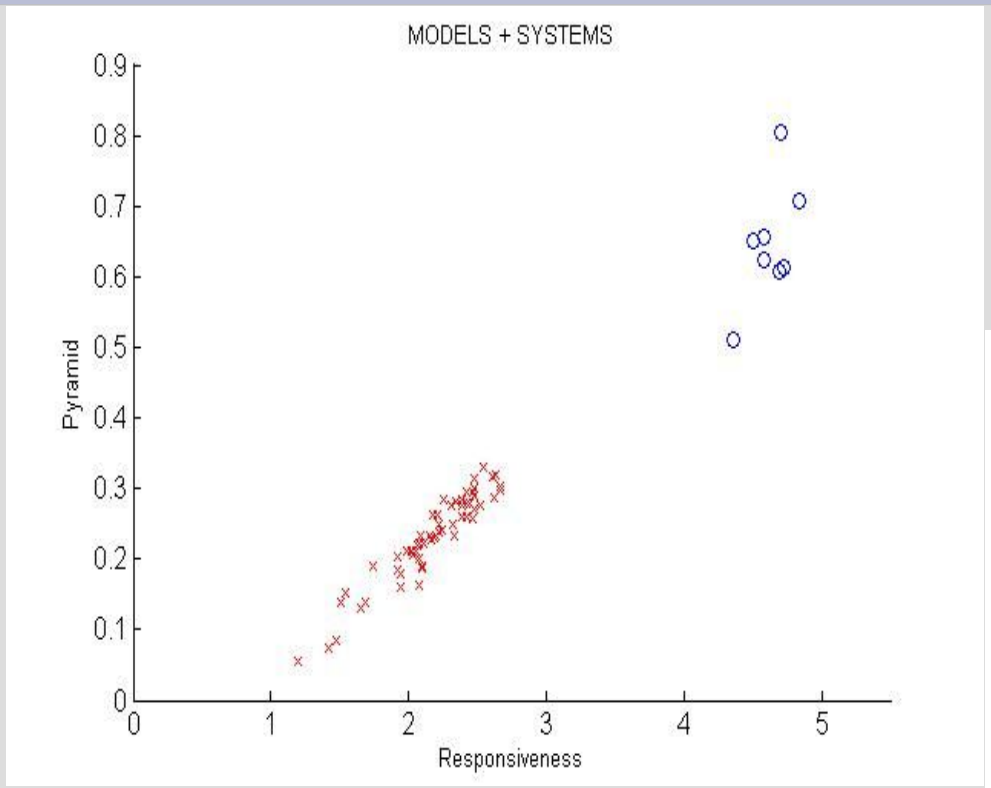
- Overall Readability – evaluation of form
- Pyramid – evaluation of content
- Overall Responsiveness – evaluation of form + content

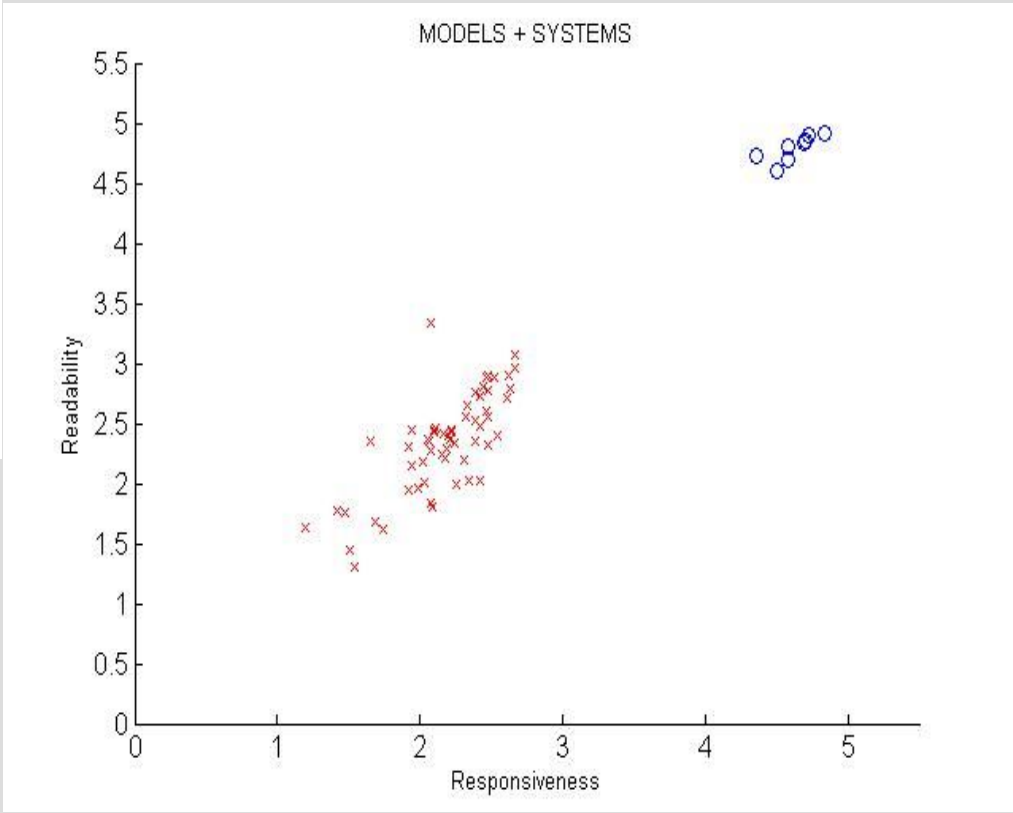
Correlation between average Responsiveness and average Readability/Pyramid

	Pearson's		Spearman's	
	models	systems	models	systems
Readability	0.778*	0.763*	0.910*	0.750*
Pyramid	0.64	0.950*	0.46	0.941*

* correlation statistically significant with $p < 0.05$

Manual Metrics - Correlation





Manual and Automatic Metrics

Correlation between Responsiveness score and ROUGE/BE

	Pearson's		Spearman's	
	models	systems	models	systems
ROUGE-2	0.725*	0.894*	0.874*	0.920*
ROUGE-SU4	0.866*	0.874*	0.898*	0.909*
BE-HM	0.656	0.911*	0.683	0.910*

Correlation between Pyramid score and ROUGE/BE

	Pearson's		Spearman's	
	models	systems	models	systems
ROUGE-2	0.276	0.946*	0.429	0.967*
ROUGE-SU4	0.457	0.928*	0.595	0.951*
BE-HM	0.423	0.949*	0.309	0.950*

* correlation statistically significant with $p < 0.05$

Conclusions

- Update summaries more difficult for automatic systems than main summaries
  - lower Overall Responsiveness
  - lower Pyramid scores
- Gap between automatic and human summaries
  - Overall Responsiveness
  - Overall Readability
  - Pyramid score
- NIST baseline best in Readability, low in content (Pyramid)

Thank you