Overview of TAC 2011 Summarization Track

Karolina Owczarzak, Hoa Trang Dang
National Institute of Standards and Technology

NIST

TAC 2010 Summarization Track

- Guided Summarization task
  - multidocument summarization
    - initial summary (100 words)
    - update summary (100 words)
  - guided by list of required aspects

- AESOP (Automatically Evaluating Summaries of Peers)
  - automatic metrics for evaluation of summary quality
  - human-crafted model summaries available
  - source documents available

NIST

Guided Summarization task

- Summarization of multiple documents on the same topic
  - initial summary:

    A 100-word summary of a set of 10 documents concerned with a single topic.
  - update summary:

    A 100-word summary of a set of further 10 documents for the same topic, with the assumption that the content of the first 10 documents is already known to the reader.
- Guided by a list of required facts (“aspects”)
  - five categories of topics
  - required aspects dependent on category
  - other important information allowed

Guided Summarization categories

1. Accidents and Natural Disasters
- 1.1 WHAT
- 1.2 WHEN
- 1.3 WHERE
- 1.4 WHY
- 1.5 WHO_AFFECTED
- 1.6 DAMAGES
- 1.7 COUNTERMEASURES

2. Attacks (Criminal/Terrorist)
- 2.1 WHAT
- 2.2 WHEN
- 2.3 WHERE
- 2.4 PERPETRATORS
- 2.5 WHY
- 2.6 WHO_AFFECTED
- 2.7 DAMAGES
- 2.8 COUNTERMEASURES

3. Health and Safety
- 3.1 WHAT
- 3.2 WHO_AFFECTED
- 3.3 HOW
- 3.4 WHY
- 3.5 COUNTERMEASURES

4. Endangered Resources
- 4.1 WHAT
- 4.2 IMPORTANCE
- 4.3 THREATS
- 4.4 COUNTERMEASURES

5. Investigations and Trials (Criminal/Legal/Other)
- 5.1 WHO
- 5.2 WHO_INVESTIGATING
- 5.3 WHY
- 5.4 CHARGES
- 5.5 PLEAD
- 5.6 SENTENCE

NIST

Guided Summarization categories

1. Accidents and Natural Disasters

D1105A Plane Crash Indonesia
D1108B Cyclone Sidr
D1110B Earthquake Sichuan
D1115C Oil Spill South Korea
D1122D Minnesota Bridge Collapse

9 topics

2. Attacks (Criminal/Terrorist)

D1116C VTech Shooting
D1123D US Embassy Greece Attack
D1126E Reporter Shoe Bush
D1139G Pirate Hijack Tanker

9 topics

3. Health and Safety

D1102A Internet Security
D1104A Pet Food Recall
D1107B China Food Safety
D1114C Heart Disease

10 topics

4. Endangered Resources

D1113C Elephants Ivory
D1120D Lake Meade Drought
D1125E Polar Bears
D1131F Endangered Coral

8 topics

5. Investigations and Trials (Criminal/Legal/Other)

D1103A Madrid Train Bombings Trial
D1117C Walter Reed Investigation
D1121D Michael Vick Dog Fight
D1128E Taylor Trial

8 topics

NIST

Guided Summarization task

- 8 NIST assessors (7 for evaluation)
- 44 topics
- 20 documents selected for each topic
  - TAC 2010 KBP Source Data: years 2007-2008, New York Times, the Associated Press, Xinhua News Agency newswires
- 20 documents divided in 2 sets
  - Set A (first 10 documents) – source text for initial summaries
  - Set B (second 10 documents) – source text for update summaries
- 4 model summaries written for each topic

NIST

Guided Summarization task

- Participants:
  - 25 teams
  - 48 runs (up to two runs per team)

	TAC 2010	TAC 2011
China	9	8
India	4	3
USA	2	6
Hong Kong	1	1
Singapore	0	1
Canada	3	3
Japan	0	1
UK	1	1
EU	1	1
Brazil	1	0
Germany	1	0

Guided Summarization task

- Baselines:
  - Baseline 1 (ID = 1): leading sentences (up to 100 words) from the most recent document
  - Baseline 2 (ID = 2): summary generated by publicly available summarizer MEAD with default settings
- All runs evaluated manually
  - Overall Responsiveness
  - Overall Readability
  - Pyramid

Guided Summarization task - Evaluation

- Overall Responsiveness

  How well does the summary respond to the information need contained in the topic statement? How good is its linguistic quality?

- Overall Readability

  How fluent and readable is the summary? Consider: grammaticality, non-redundancy, referential clarity, focus, structure, coherence.

Very Poor Poor Barely Acceptable Good Very Good
1.....2.....3.....4.....5

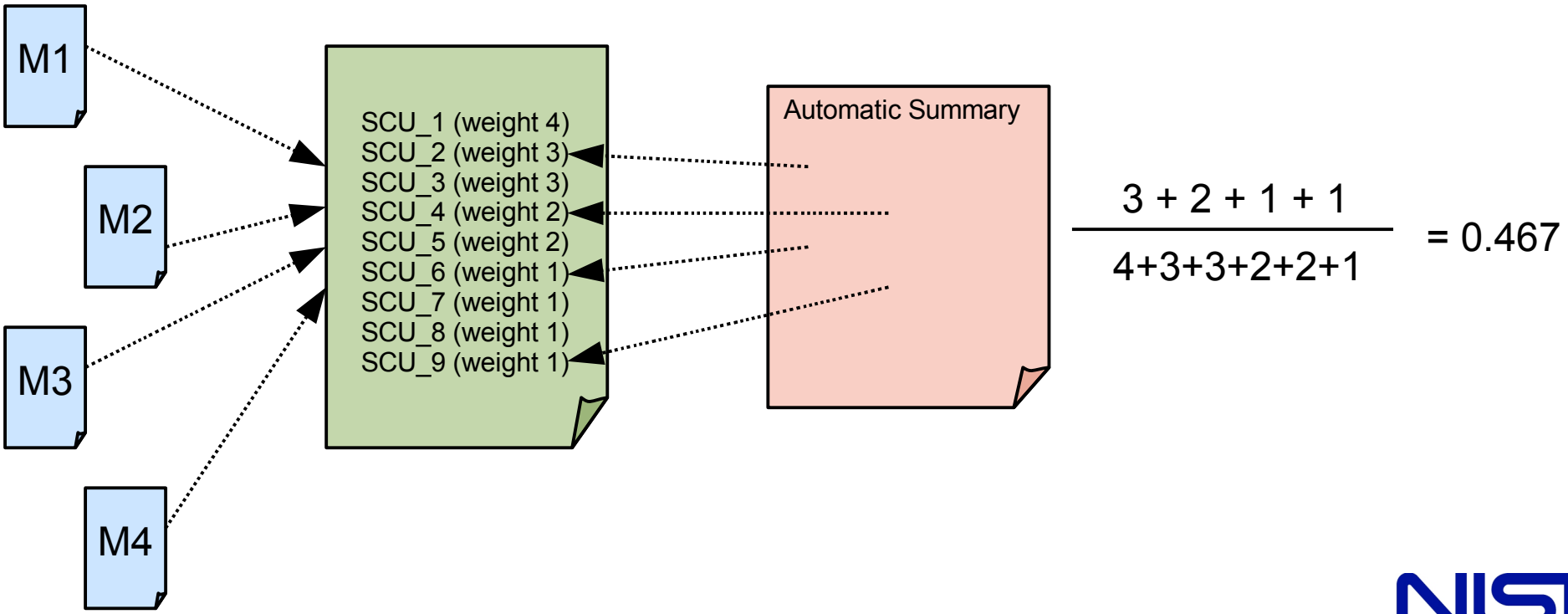
- System score = mean score of all its summaries
- System ranking
  - ANOVA
  - multiple comparison (Tukey's honestly significant difference criterion)

NIST

Guided Summarization task - Evaluation

- Pyramid (Passonneau et al., 2005)

$$\text{score} = \frac{\text{total weight of all SCUs present in the candidate}}{\text{total SCU weight possible for average-length summary}}$$



Evaluation - Responsiveness

ID	Score		ID	Score	
D	4.9545	A	G	4.9091	A
C	4.9545	A	H	4.8636	A
H	4.9091	A	D	4.7727	A
A	4.8182	A	A	4.7727	A
E	4.7727	A	C	4.6818	A
G	4.7273	A	E	4.5455	A
B	4.7273	A	B	4.5000	A
F	4.6818	A	F	4.3182	A
CLASSY2	3.1591	B	SIEL_IIITH2	2.5909	B
PKUTM2	3.1364	BC	seme11	2.5682	BC
TJU_Summary1	3.1136	BC	pris1	2.5455	BCD
pris1	3.0909	BC	CLASSY2	2.5455	BCD
pris2	3.0909	BC	IIScSum1	2.5227	BCD
NUS2	3.0909	BC	PolyCom1	2.5227	BCD
seme11	3.0682	BCD	NUS2	2.5000	BCD
NUS1	3.0682	BCD	SIEL_IIITH1	2.5000	BCD
SIEL_IIITH1	3.0455	BCD	seme12	2.4773	BCD
BLLIP2	3.0227	BCD	PKUTM2	2.4773	BCD
(Baseline2	*2.8409)*		*(Baseline2*	*2.1136)*	
(Baseline1	*2.5000)*		*(Baseline1*	*2.0909)*	

models

Initial summaries

Update summaries

Evaluation - Readability

ID	Score		models	ID	Score	
E	5.0000	A		H	5.0000	A
D	5.0000	A		C	4.9545	A
C	5.0000	A		G	4.9091	A
H	4.9545	A		E	4.9091	A
A	4.8636	A		B	4.9091	A
B	4.8182	A		A	4.9091	A
G	4.7273	A		D	4.8636	A
F	4.5909	AB		F	4.7273	A
pris1	3.7500	BC		*Baseline1*	*3.4545*	*B*
pris2	3.5227	CD		pris1	3.3409	BC
seme11	3.5000	CD		CLASSY2	3.3409	BC
JRC1	3.4545	CDE		UW_20112	3.3409	BC
PKUTM2	3.4318	CDEF		PKUTM2	3.2727	BCD
CLASSY2	3.3409	CDEFG		JRC1	3.2500	BCDE
Baseline1	*3.2045*	*CDEFGH*		seme11	3.2273	BCDEF
seme12	3.1818	CDEFGH		uOttawa2	3.0909	BCDEF
uOttawa1	3.1364	CDEFGH		seme12	3.0682	BCDEF
CLASSY1	3.1364	CDEFGH		CLASSY1	3.0682	BCDEF
(Baseline2	*2.8182)*			*(Baseline2*	*2.8409)*	

Initial summaries

Update summaries

Evaluation - Pyramid

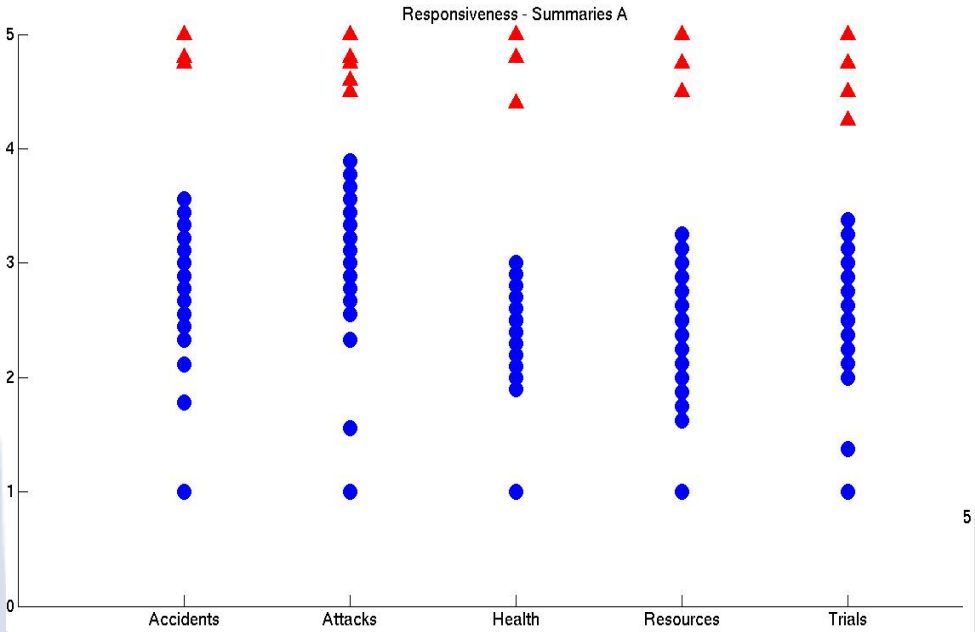
ID	Score		ID	Score	
G	0.88791	A	D	0.82305	A
D	0.83759	A	G	0.72818	AB
H	0.79959	A	H	0.71909	AB
B	0.78082	A	A	0.66350	AB
A	0.77068	A	F	0.62391	AB
C	0.75205	A	E	0.61545	B
E	0.72168	A	C	0.56541	B
F	0.70491	A	B	0.55364	B
PKUTM2	0.47077	B	IISCSum1	0.34645	C
NUS1	0.46836	BC	ICTCAS2	0.34641	C
NUS2	0.46223	BCD	NUS1	0.34270	C
PolyCom1	0.44727	BCDE	CLASSY2	0.33748	C
BLLIP1	0.44084	BCDE	SIEL_IIITH2	0.33680	C
seme11	0.43741	BCDE	TJU_GSummary2	0.33327	C
PolyCom2	0.43741	BCDE	NUS2	0.33275	C
BLLIP2	0.43734	BCDE	PolyCom1	0.33184	C
pris1	0.43573	BCDE	ICTCAS1	0.33014	C
CLASSY2	0.43559	BCDE	seme11	0.32575	C
(Baseline2	*0.35743)*		*(Baseline2*	*0.27980)*	
(Baseline1	*0.29989)*		*(Baseline1*	*0.23425)*	

models

Initial summaries

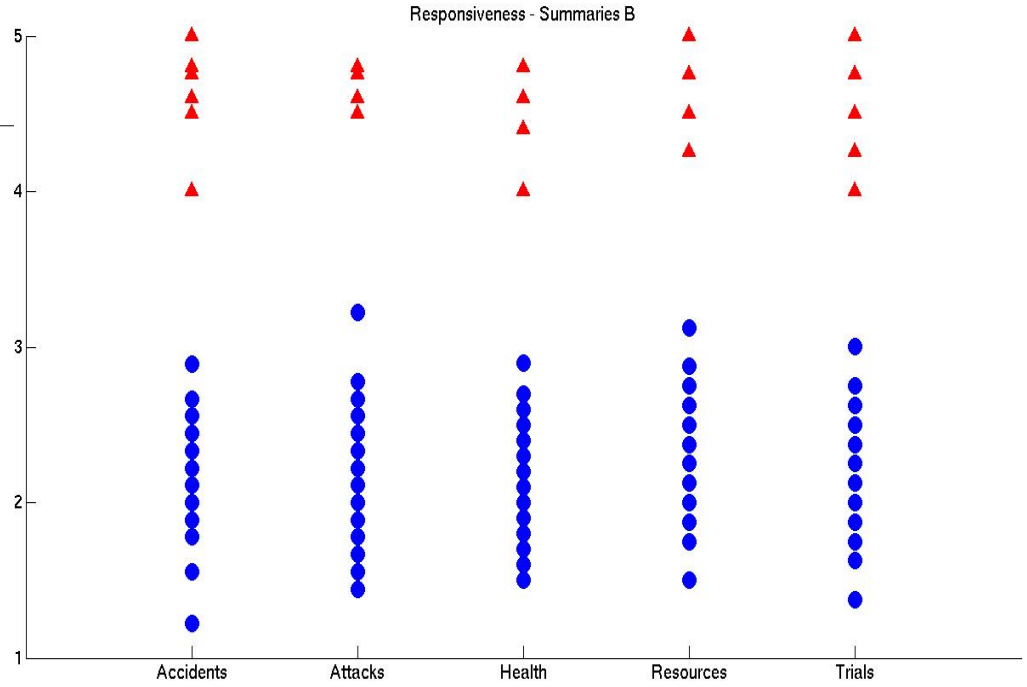
Update summaries

Evaluation – Responsiveness Averages



Summaries A

CatID	Human		CatID	Automatic	
Acc	4.944	A	Att	3.018	A
Att	4.833	AB	Acc	2.916	A
Hea	4.800	ABC	Tri	2.698	B
Res	4.781	ABCD	Hea	2.400	C
Tri	4.719	ABCD	Res	2.398	C

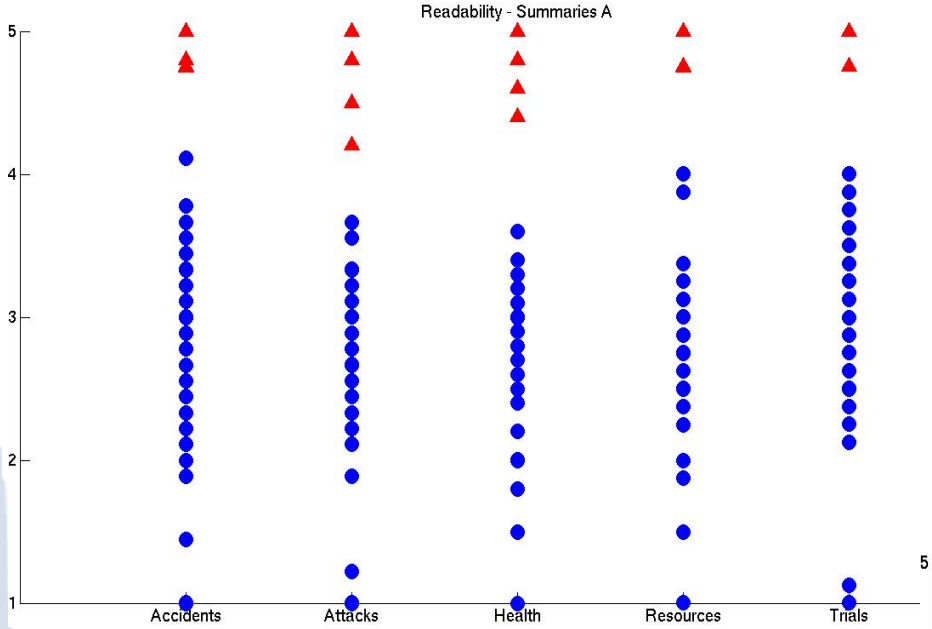


Summaries B

CatID	Human		CatID	Automatic	
Res	4.719	A	Res	2.350	A
Acc	4.694	AB*	Tri	2.260	AB*
Att	4.694	ABC*	Hea	2.242	ABC*
Hea	4.625	ABCD*	Att	2.189	BCD*
Tri	4.625	ABCD	Acc	2.129	BCD*

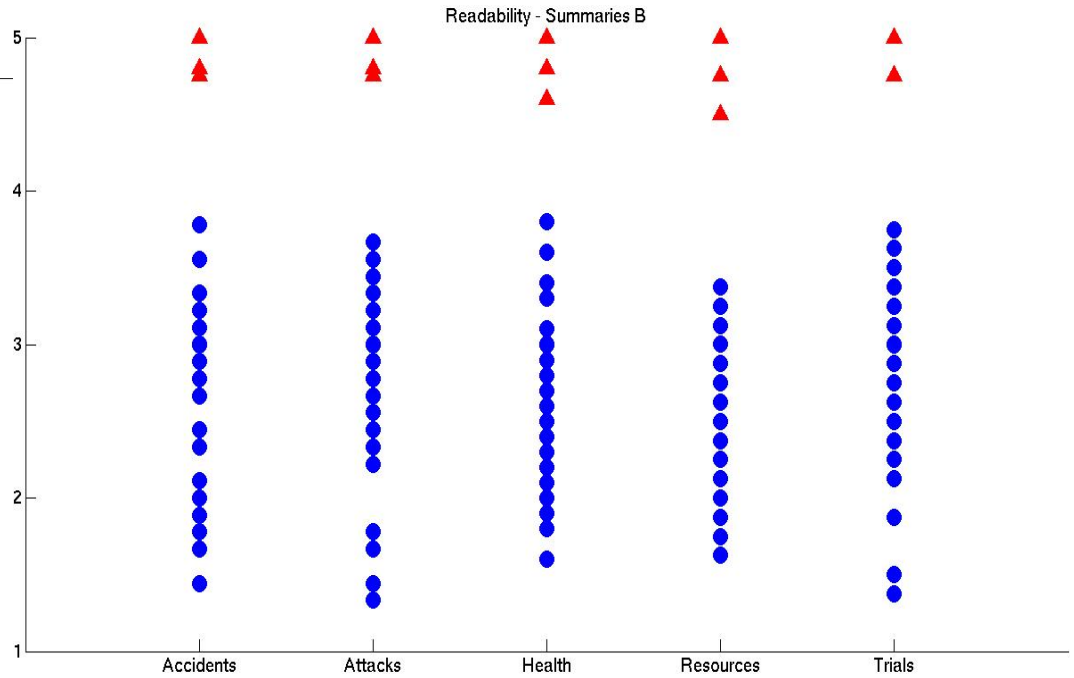
* significant drop from the initial score

Evaluation – Readability Averages



Summaries A

CatID	Human		CatID	Automatic	
Tri	4.938	A	Acc	2.853	A
Res	4.906	AB	Tri	2.850	AB
Acc	4.889	ABC	Hea	2.740	ABC
Hea	4.825	ABCD	Att	2.691	ABCD
Att	4.806	ABCD	Res	2.672	ABCD

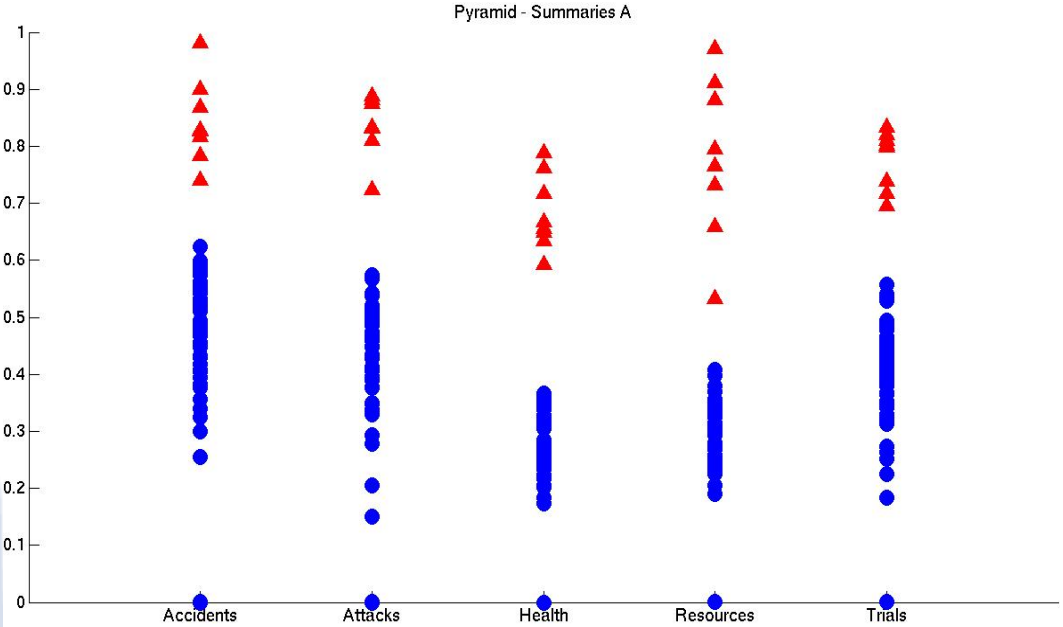


Summaries B

CatID	Human		CatID	Automatic	
Acc	4.917	A	Acc	2.804	A
Att	4.917	AB*	Tri	2.788	AB
Hea	4.900	ABC*	Att	2.733	ABC
Res	4.875	ABCD	Hea	2.698	ABCD
Tri	4.875	ABCD	Res	2.670	ABCD

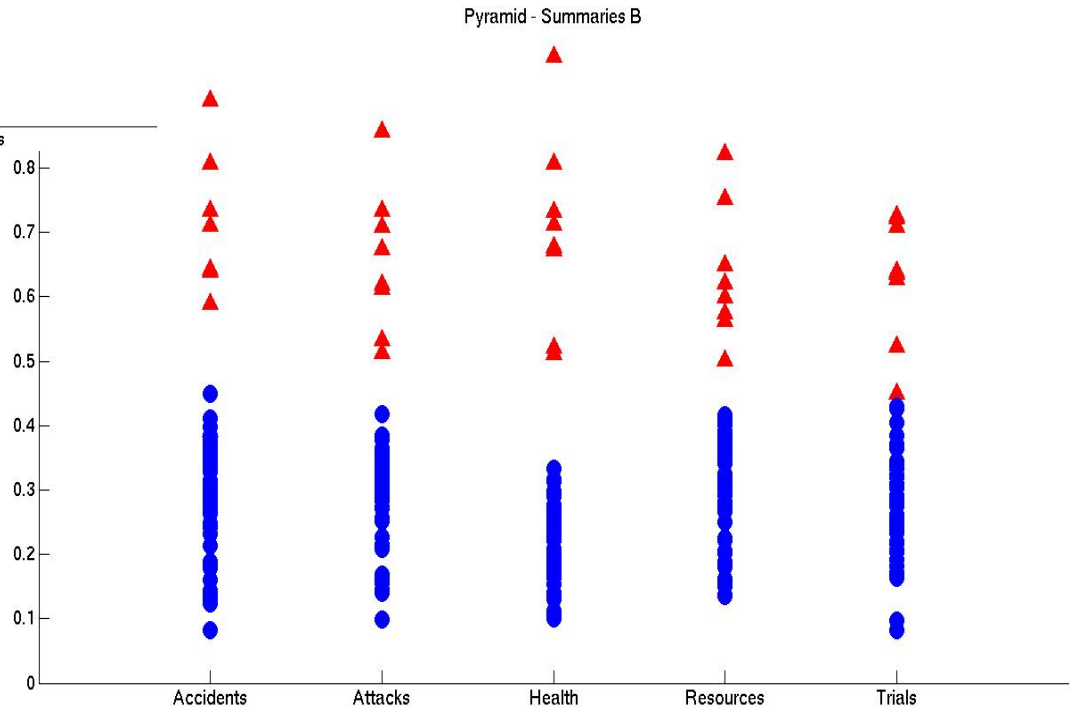
* significant increase from the initial score (cf. the charts)

Evaluation – Pyramid Averages



Summaries A

CatID	Human		CatID	Automatic	
Acc	0.848	A	Acc	0.468	A
Att	0.831	A	Att	0.420	B
Res	0.781	A	Tri	0.389	B
Tri	0.776	A	Res	0.286	C
Hea	0.683	B	Hea	0.278	C



Summaries B

CatID	Human		CatID	Automatic	
Hea	0.700	A	Att	0.293	A*
Acc	0.683	AB*	Res	0.286	AB
Att	0.650	ABC*	Acc	0.277	ABC*
Res	0.635	ABCD*	Tri	0.270	ABC*
Tri	0.628	ABCD*	Hea	0.217	D*

* significant drop from the initial score

Measuring redundancy

- Evaluating update summaries against joined Pyramid A+B

Number of SCUs from Pyramid A in summaries B

	Accidents	Attacks	Health	Resources	Trials
automatic summarizers	4	6	2	2	4
models (true)	4 (2)	5 (2)	1 (0)	3 (2)	3 (1)

Guided Summarization task - Conclusions

- Gap between models and automatic summaries
- Many automatic summarizers better than baselines (except Readability)
- Automatic summarizers:
  - lower avg content scores in Health, Resources
  - lower avg content scores in update part
- Human summarizers:
  - slightly lower avg Responsiveness in update part
  - lower avg Pyramid scores in update part (= less content overlap)

AESOP task

- Goal: emulate Pyramid, Responsiveness, Readability
- Test data:
  - 51 automatic summarizers
  - 8 human summarizers (4 models per topic)
  - 44 topics (A & B): summaries, source documents, topic titles
- Participants
  - 7 teams
  - 22 metrics (up to 4 runs per team)
- Baselines:
  - ROUGE-2: matching bigrams, stemmed (Lin, 2004)
  - ROUGE-SU4: matching bigrams with skip distance up to 4 words, stemmed (Lin, 2004)
  - BE-HM: head-modifier pairs, stemmed (Hovy et al., 2005)

NIST

AESOP task

- Use of resources:
  - model summaries: 17 metrics
  - source documents: 6 metrics
  - topic titles used: 4 metrics
- Conditions:
  - AllPeers: models + automatic summaries

    Can automatic metrics distinguish between human and automatic summaries?
  - NoModels: only automatic summaries, model summaries as reference

    Can automatic metrics accurately evaluate the quality of automatic summaries?
  - Summarizer-level: ranking of summarizers
  - Summary-level: ranking of individual summaries

NIST

AESOP task - Evaluation

- Overall Responsiveness
  - content relevance to topic and aspects
  - linguistic quality
- Overall Readability
  - linguistic quality, focus, structure, non-redundancy
- Pyramid
  - content similarity between candidate and reference summaries
  - guided summarization = more similar models

initial	2008	2009	2010	2011
human	0.66	0.68	0.78	0.78
automatic	0.26	0.26	0.30	0.37

update	2008	2009	2010	2011
human	0.63	0.60	0.67	0.66
automatic	0.20	0.20	0.20	0.27

Macro-average Pyramid scores for years 2008 - 2010

NIST

AESOP task - Evaluation

- Summarizer-level and summary-level correlations
- Correlations (Pearson, Spearman, Kendall) with:
  - Overall Responsiveness
  - Overall Readability
  - Pyramid
- Discriminative power

AESOP metric

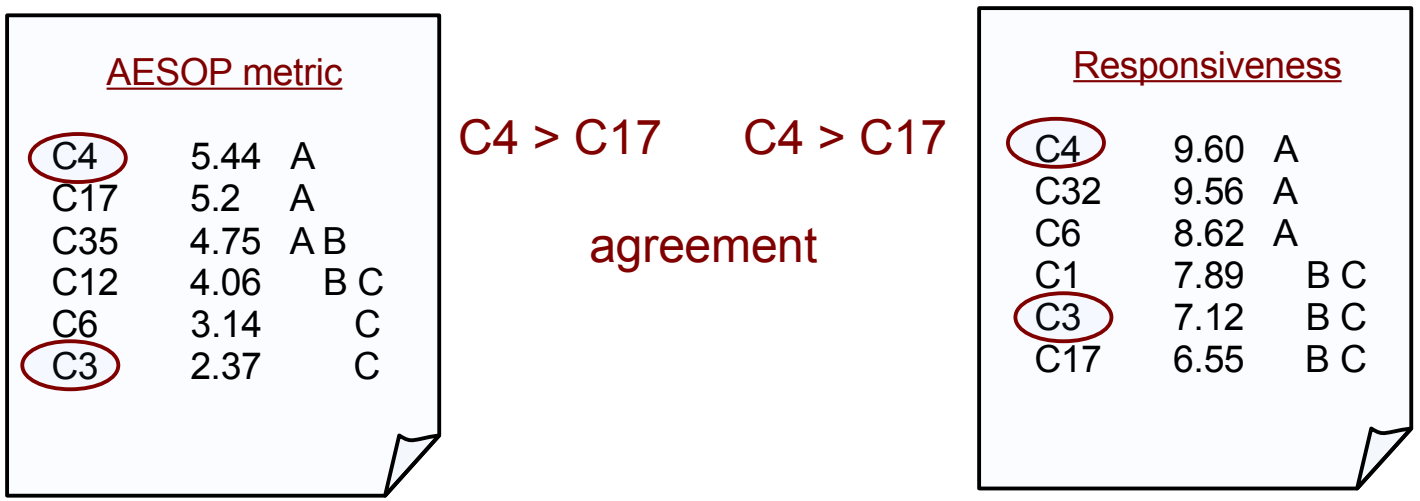
C4	5.44	A
C17	5.2	A
C35	4.75	A B
C12	4.06	B C
C6	3.14	C
C3	2.37	C

Responsiveness

C4	9.60	A
C32	9.56	A
C6	8.62	A
C1	7.89	B C
C3	7.12	B C
C17	6.55	B C

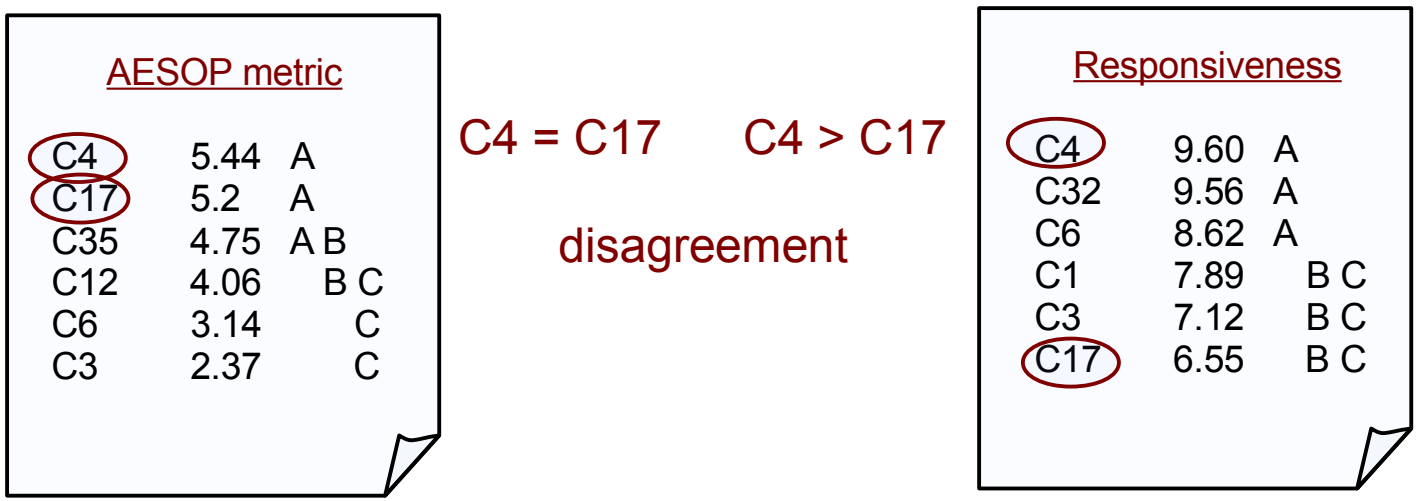
AESOP task - Evaluation

- Summarizer-level and summary-level correlations
- Correlations (Pearson, Spearman, Kendall) with:
  - Overall Responsiveness
  - Overall Readability
  - Pyramid
- Discriminative power



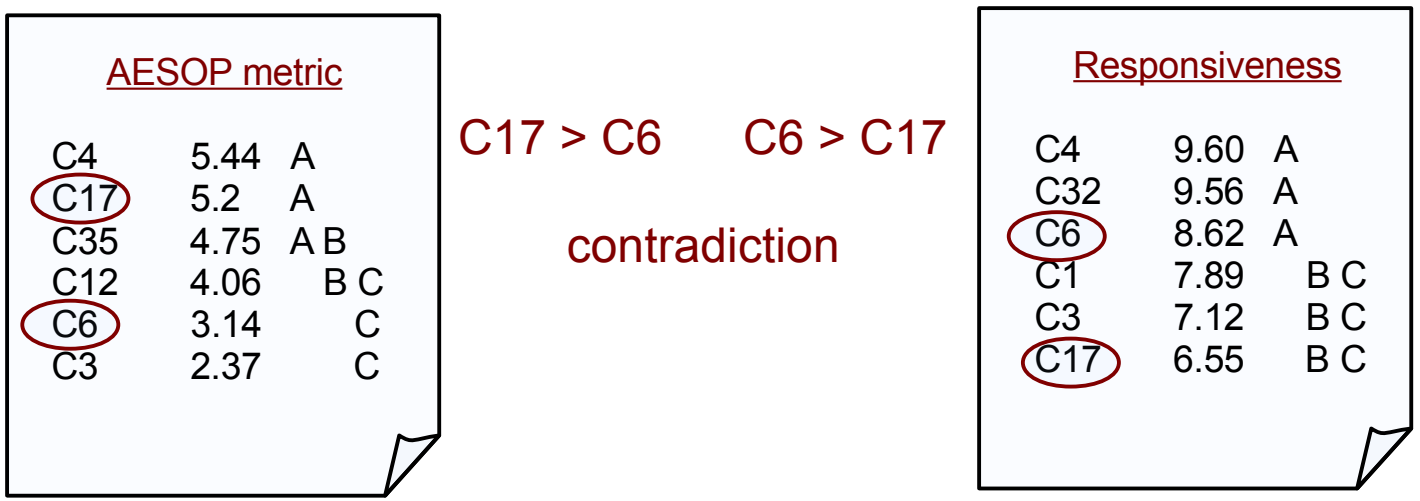
AESOP task - Evaluation

- Summarizer-level and summary-level correlations
- Correlations (Pearson, Spearman, Kendall) with:
  - Overall Responsiveness
  - Overall Readability
  - Pyramid
- Discriminative power



AESOP task - Evaluation

- Summarizer-level and summary-level correlations
- Correlations (Pearson, Spearman, Kendall) with:
  - Overall Responsiveness
  - Overall Readability
  - Pyramid
- Discriminative power



Pearson's r – NoModels, ranking systems

Pyramid

ROUGE-SU4	*0.981*
DemokritosGR1	0.974
CLASSY4	0.968
PKUTM1	0.968
catolicasc1	0.967
CLASSY2	0.967
C_S_IIITH3	0.965
DemokritosGR2	0.964
PKUTM4	0.962
PKUTM3	0.962

CLASSY4	0.911
BE-HM	*0.906*
PKUTM3	0.904
ROUGE-2	*0.903*
CLASSY2	0.900
CLASSY1	0.898
CLASSY3	0.890
DemokritosGR2	0.885
ROUGE-SU4	*0.885*
C_S_IIITH3	0.884

Readability

catolicasc1	0.819
DemokritosGR1	0.794
DemokritosGR2	0.791
CLASSY4	0.784
ROUGE-SU4	*0.784*
CLASSY1	0.778
C_S_IIITH1	0.777
C_S_IIITH2	0.776
CLASSY2	0.774
C_S_IIITH4	0.773

catolicasc1	0.742
CLASSY3	0.705
CLASSY4	0.683
ROUGE-SU4	*0.672*
DemokritosGR2	0.670
PKUTM3	0.662
ROUGE-2	*0.658*
DemokritosGR1	0.644
CLASSY2	0.620
C_S_IIITH4	0.620

Responsiveness

ROUGE-SU4	*0.954*
CLASSY4	0.951
CLASSY2	0.951
catolicasc1	0.950
CLASSY1	0.949
DemokritosGR2	0.948
DemokritosGR1	0.947
PKUTM3	0.943
ROUGE-2	*0.942*
PKUTM1	0.936

CLASSY4	0.927
PKUTM3	0.919
CLASSY3	0.919
ROUGE-2	*0.917*
ROUGE-SU4	*0.912*
CLASSY2	0.903
CLASSY1	0.903
DemokritosGR2	0.891
C_S_IIITH3	0.885
BE-HM	*0.876*

NIST

Pearson's r – AllPeers, ranking systems

Pyramid

C_S_IIITH1	0.975
catolicasc1	0.974
C_S_IIITH2	0.956
DemokritosGR2	0.951
C_S_IIITH4	0.950
CLASSY1	0.945
CLASSY2	0.945
CLASSY3	0.909
CLASSY4	0.853
DemokritosGR1	0.842
C_S_IIITH3	0.786

CLASSY3	0.953
CLASSY4	0.953
catolicasc1	0.950
CLASSY2	0.944
C_S_IIITH1	0.938
CLASSY1	0.936
DemokritosGR2	0.933
C_S_IIITH2	0.882
C_S_IIITH4	0.865
DemokritosGR1	0.824
ROUGE-2	*0.775*

Readability

catolicasc1	0.926
C_S_IIITH1	0.906
DemokritosGR2	0.906
CLASSY1	0.903
CLASSY2	0.903
C_S_IIITH2	0.894
C_S_IIITH4	0.884
CLASSY3	0.844
CLASSY4	0.774
DemokritosGR1	0.770
C_S_IIITH3	0.711

catolicasc1	0.934
CLASSY2	0.915
CLASSY1	0.915
CLASSY3	0.907
DemokritosGR2	0.895
CLASSY4	0.887
C_S_IIITH1	0.868
C_S_IIITH2	0.837
C_S_IIITH4	0.822
DemokritosGR1	0.761
ROUGE-2	*0.712*

Responsiveness

catolicasc1	0.972
C_S_IIITH1	0.965
DemokritosGR2	0.963
CLASSY1	0.948
CLASSY2	0.948
C_S_IIITH2	0.937
C_S_IIITH4	0.929
CLASSY3	0.899
CLASSY4	0.830
DemokritosGR1	0.815
BE-HM	*0.752*

DemokritosGR2	0.975
catolicasc1	0.974
CLASSY1	0.965
CLASSY2	0.963
CLASSY3	0.961
CLASSY4	0.949
C_S_IIITH1	0.937
C_S_IIITH2	0.880
C_S_IIITH4	0.859
DemokritosGR1	0.774
ROUGE-2	*0.717*

NIST

Pearson's r – NoModels, ranking summaries

Pyramid

DemokritosGR1	0.752
ROUGE-SU4	*0.736*
PKUTM4	0.732
PKUTM1	0.732
PKUTM2	0.726
CLASSY4	0.721
CLASSY2	0.721
PKUTM3	0.710
ROUGE-2	*0.709*
CLASSY3	0.705

DemokritosGR1	0.520
BE-HM	*0.512*
DemokritosGR2	0.505
ROUGE-SU4	*0.499*
catolicasc1	0.482
PKUTM3	0.472
ROUGE-2	*0.465*
CLASSY4	0.449
CLASSY3	0.420
C_S_IIITH1	0.407

Readability

catolicasc1	0.511
DemokritosGR2	0.497
DemokritosGR1	0.496
CLASSY4	0.467
C_S_IIITH1	0.466
ROUGE-SU4	*0.459*
PKUTM1	0.451
PKUTM4	0.448
CLASSY2	0.445
PKUTM2	0.440

catolicasc1	0.361
DemokritosGR1	0.321
DemokritosGR2	0.320
C_S_IIITH1	0.318
ROUGE-SU4	*0.304*
uOttawa2	0.287
uOttawa3	0.280
PKUTM3	0.268
CLASSY3	0.263
ROUGE-2	*0.261*

Responsiveness

average correlations per assessor = avoid inter-rater variance

DemokritosGR1	0.632
DemokritosGR2	0.625
ROUGE-SU4	*0.614*
CLASSY4	0.611
catolicasc1	0.608
PKUTM1	0.607
CLASSY2	0.606
PKUTM4	0.604
CLASSY1	0.594
PKUTM2	0.593

DemokritosGR1	0.476
DemokritosGR2	0.470
ROUGE-SU4	*0.445*
BE-HM	*0.432*
catolicasc1	0.425
PKUTM3	0.406
ROUGE-2	*0.399*
CLASSY4	0.395
CLASSY3	0.387
C_S_IIITH1	0.380

NIST

Rater consistency

- Inter-rater agreement vs. intra-rater agreement (rater consistency)
- Identical summaries in Guided task (variations of same system):
  - 417 pairs of summaries
  - around 60 pairs per assessor

Krippendorff's alpha for interval values

Assessor ID	Pyramid	Responsiveness	Readability
A	0.93	0.75	0.80
C	0.89	0.49	0.64
D	0.97	0.88	0.87
E	0.91	0.71	0.52
F	0.87	0.73	0.65
G	0.98	0.93	0.87
H	0.95	0.95	0.77

Pearson's r – NoModels, ranking summaries

Pyramid

DemokritosGR1	0.781
ROUGE-SU4	*0.754*
PKUTM1	0.744
PKUTM4	0.743
CLASSY4	0.741
DemokritosGR2	0.739
catolicasc1	0.739
CLASSY2	0.738
PKUTM2	0.735
PKUTM3	0.720

BE-HM	*0.569*
catolicasc1	0.557
ROUGE-SU4	*0.554*
DemokritosGR1	0.553
DemokritosGR2	0.527
PKUTM3	0.516
ROUGE-2	*0.508*
CLASSY4	0.492
uOttawa2	0.472
CLASSY3	0.444

Readability

catolicasc1	0.559
DemokritosGR2	0.552
DemokritosGR1	0.547
C_S_IIITH1	0.535
CLASSY4	0.513
ROUGE-SU4	*0.504*
C_S_IIITH2	0.500
C_S_IIITH4	0.490
PKUTM1	0.489
PKUTM4	0.488

catolicasc1	0.380
C_S_IIITH1	0.325
DemokritosGR1	0.323
DemokritosGR2	0.322
ROUGE-SU4	*0.308*
C_S_IIITH4	0.293
uOttawa2	0.285
C_S_IIITH2	0.281
uOttawa3	0.279
CLASSY3	0.268

Responsiveness

average correlations per assessor; exclude low-consistency C,E,F

DemokritosGR2	0.670
DemokritosGR1	0.669
ROUGE-SU4	*0.652*
CLASSY4	0.652
PKUTM1	0.644
CLASSY2	0.644
PKUTM4	0.644
catolicasc1	0.642
PKUTM2	0.637
CLASSY1	0.626

ROUGE-SU4	*0.502*
DemokritosGR1	0.499
DemokritosGR2	0.493
catolicasc1	0.480
BE-HM	*0.479*
PKUTM3	0.451
ROUGE-2	*0.441*
CLASSY4	0.436
CLASSY3	0.418
uOttawa2	0.414

Evaluation – Discriminative power

Initial summaries				Update summaries			
ID	difference (max 408)	no difference (max 0)	contradiction	ID	difference (max 408)	no difference (max 0)	contradiction
catolicasc1	408	0	0	catolicasc1	408	0	0
C_S_IIITH2	408	0	0	C_S_IIITH2	408	0	0
DemokritosGR2	408	0	0	DemokritosGR2	408	0	0
C_S_IIITH4	408	0	0	C_S_IIITH4	408	0	0
C_S_IIITH1	408	0	0				
ROUGE-SU4	*102*	*0*	*0*	*ROUGE-SU4*	*132*	*0*	*0*
BE-HM	*80*	*0*	*0*	*ROUGE-2*	*114*	*0*	*0*
ROUGE-2	*78*	*0*	*0*	*BE-HM*	*75*	*0*	*0*

Finding significant differences between human and automatic summarizers – AESOP metrics vs. Pyramid/Responsiveness

Initial summaries				Update summaries			
ID	difference (max 407)	no difference (max 1)	contradiction	ID	difference (max 408)	no difference (max 0)	contradiction
catolicasc1	407	0	0	catolicasc1	408	0	0
C_S_IIITH2	407	0	0	C_S_IIITH2	408	0	0
DemokritosGR2	407	0	0	DemokritosGR2	408	0	0
C_S_IIITH4	407	0	0	C_S_IIITH1	408	0	0
C_S_IIITH1	407	0	0				
ROUGE-SU4	*102*	*0*	*0*	*BE-HM*	*132*	*0*	*0*
BE-HM	*80*	*0*	*0*	*ROUGE-SU4*	*114*	*0*	*0*
ROUGE-2	*78*	*0*	*0*	*ROUGE-2*	*75*	*0*	*0*

Finding significant differences between human and automatic summarizers – AESOP metrics vs. Readability

NIST

Evaluation – Discriminative power

Initial summaries				Update summaries			
ID	difference (max 239)	no difference (max 1036)	contradiction	ID	difference (max 187)	no difference (max 1088)	contradiction
CLASSY4	236	752	0	*ROUGE-SU4*	*157*	*953*	*0*
DemokritosGR1	236	825	0	uOttawa3	154	895	0
CLASSY2	236	762	0	PKUTM3	151	993	0
PKUTM3	235	837	0	*ROUGE-2*	*150*	*998*	*0*
DemokritosGR2	235	790	0	C S IIITH1	150	953	0
ROUGE-2	*235*	*835*	*0*	uOttawa1	146	623	0
ROUGE-SU4	*235*	*809*	*0*	catolicasc1	145	951	0
BE-HM	*220*	*891*	*0*	*BE-HM*	*143*	*1036*	*0*

Finding significant differences between automatic summarizers – AESOP metrics vs. Pyramid

Initial summaries				Update summaries			
ID	difference (max 221)	no difference (max 1054)	contradiction	ID	difference (max 128)	no difference (max 1147)	contradiction
DemokritosGR2	218	791	0	*ROUGE-SU4*	*125*	*980*	*0*
CLASSY4	216	750	0	C S IIITH1	122	984	0
DemokritosGR1	216	823	0	PKUTM3	121	1022	0
catolicasc1	216	818	0	catolicasc1	121	986	0
CLASSY2	216	760	0	DemokritosGR2	120	1063	0
ROUGE-2	*214*	*832*	*0*	ROUGE-2	120	1027	0
ROUGE-SU4	*213*	*805*	*0*	CLASSY4	115	1075	0
BE-HM	*201*	*890*	*0*	*BE-HM*	*110*	*1062*	*0*

Finding significant differences between automatic summarizers – AESOP metrics vs. Responsiveness

NIST

Evaluation – Discriminative power

Initial summaries

ID	difference (max 414)	no difference (max 861)	contradiction
catolicasc1	279	688	0
CLASSY4	273	614	0
CLASSY3	279	649	0
CLASSY2	270	621	0
CLASSY1	264	674	0
ROUGE-SU4	*259*	*658*	*0*
ROUGE-2	*249*	*674*	*0*
BE-HM	*217*	*713*	*0*

Update summaries

ID	difference (max 325)	no difference (max 950)	contradiction
catolicasc1	227	985	*0*
ROUGE-SU4	*176*	*835*	*0*
uOttawa1	175	525	0
PKUTM3	164	868	*0*
ROUGE-2	*163*	*873*	*0*
uOttawa3	157	763	0
CLASSY3	154	914	0
BE-HM	*124*	*879*	*0*

Finding significant differences between automatic summarizers – AESOP metrics vs. Readability

AESOP task - Conclusions

- Correlations with manual metrics
  - higher with content measures (Pyramid, Responsiveness)
  - lower with Readability
- Summary-level correlations
  - higher than expected, esp. after removing low-consistency assessors
  - room for improvement
- Discriminative power
  - some AESOP metrics perfectly match manual metrics in human vs. auto summarizers; much better than baselines
  - very high agreements in distinguishing among automatic summarizers only
  - lower agreements for Responsiveness

NIST

Thank you