Overview of TAC 2010 Summarization Track

AESOP Task

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

AESOP task

- Goal: emulate Pyramid and/or Responsiveness
- Test data:
  - 43 automatic summarizers
  - 8 human summarizers (4 models per topic)
  - 46 topics (A & B): summaries, source documents, topic titles
- Participants
  - 9 teams
  - 24 metrics (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: 23 metrics
  - source documents: 1 metric

- 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?

AESOP task - Evaluation

- Overall Responsiveness
  - content relevance to topic and aspects
  - linguistic quality
- Pyramid
  - content similarity between candidate and reference summaries
  - guided summarization = more similar models

2008	initial	update
human	0.66	0.63
automatic	0.26	0.20

2009	initial	update
human	0.68	0.60
automatic	0.26	0.20

2010	initial	update
human	0.78	0.67
automatic	0.30	0.20

Macro-average Pyramid scores for years 2008 - 2010

AESOP task - Evaluation

- Correlations (Pearson, Spearman, Kendall) with:
  - Overall Responsiveness
  - 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

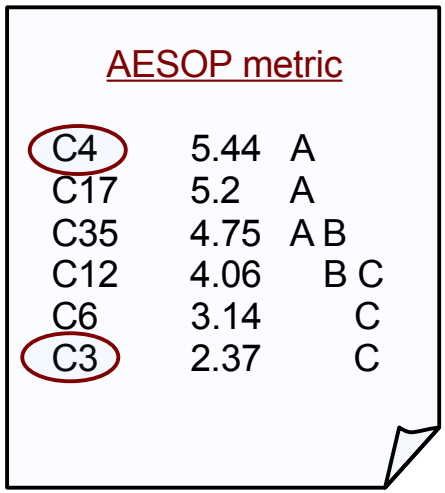
Responsiveness

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

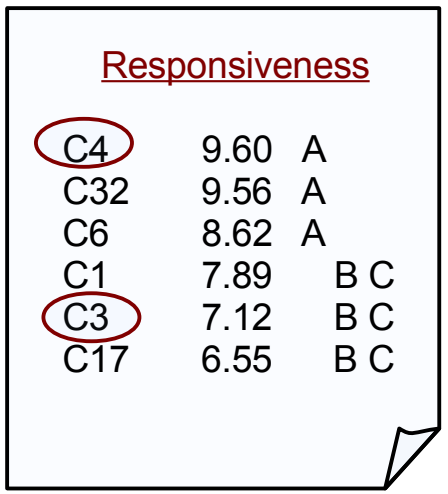
- Correlations (Pearson, Spearman, Kendall) with:
  - Overall Responsiveness
  - Pyramid

- Discriminative power



C4 > C3

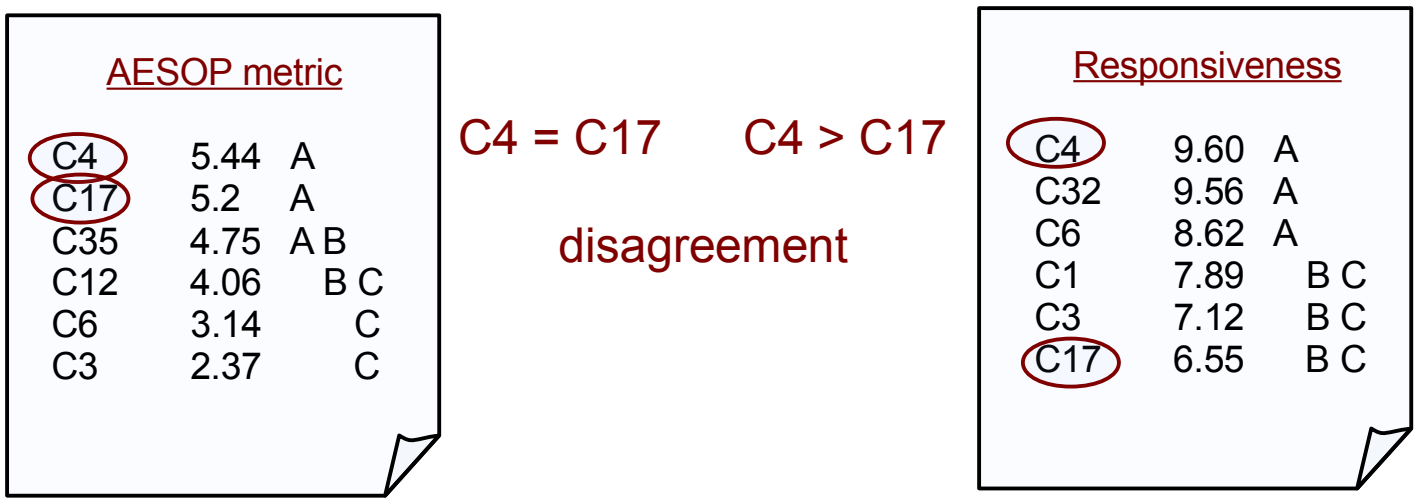
agreement



AESOP task - Evaluation

- Correlations (Pearson, Spearman, Kendall) with:
  - Overall Responsiveness
  - Pyramid

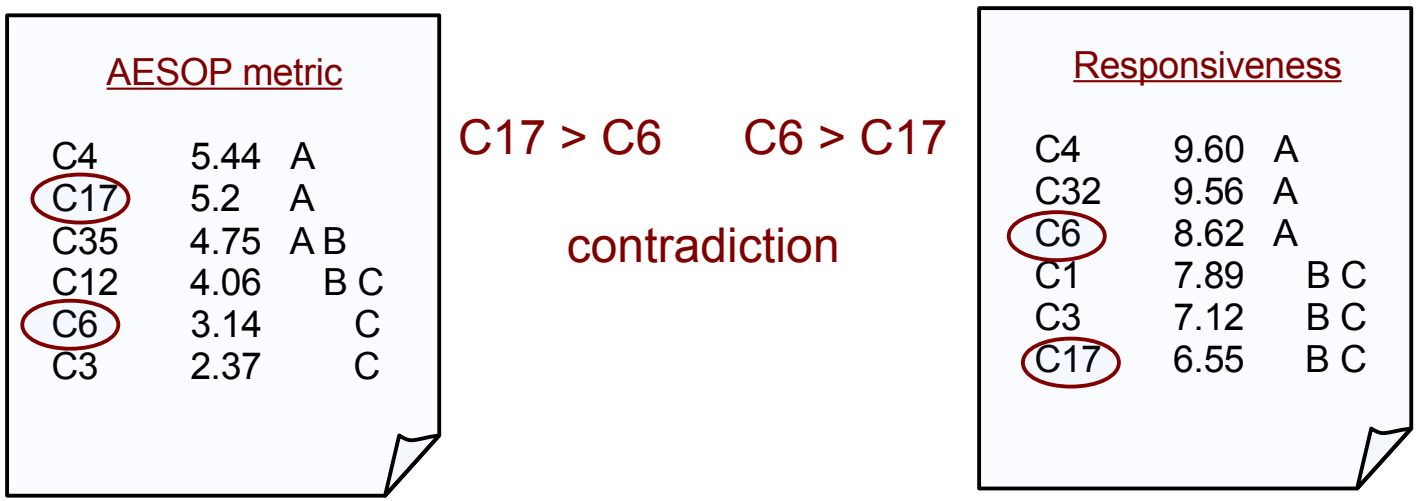
- Discriminative power



AESOP task - Evaluation

- Correlations (Pearson, Spearman, Kendall) with:
  - Overall Responsiveness
  - Pyramid

- Discriminative power



Evaluation – Correlations NoModels

Pyramid

ROUGE-2	*0.978*
CLASSY1	0.976
CLASSY2	0.972
DemokritosGR2	0.970
ROUGE-SU4	*0.968*
BE-HM	0.965
CLASSY3	0.961
CLASSY4	0.961
ICL_SUM2	0.960
DemokritosGR3	0.951

Initial summaries (NoModels)

Update summaries (NoModels)

CLASSY1	0.964
ROUGE-2	*0.963*
CLASSY4	0.959
CLASSY2	0.958
DemokritosGR2	0.956
BE-HM	*0.953*
ICL_SUM2	0.918
ROUGE-SU4	*0.910*
PolyU3	0.906
uOttawa3	0.905

Responsiveness

Initial summaries (NoModels)

CLASSY1	0.979
CLASSY2	0.975
ROUGE-2	*0.967*
CLASSY3	0.961
ROUGE-SU4	*0.955*
CLASSY4	0.954
DemokritosGR2	0.949
BE-HM	*0.943*
DemokritosGR3	0.932
ICL SUM2	0.929

CLASSY4	0.960
CLASSY1	0.959
CLASSY2	0.956
ROUGE-2	*0.953*
DemokritosGR2	0.933
BE-HM	0.928
ROUGE-SU4	*0.899*
CLASSY3	0.899
uOttawa3	0.889
DemokritosGR3	0.888

Update summaries (NoModels)

NIST

Evaluation – Correlations AllPeers

Pyramid

NIRAJIIITH1	0.975
Univille1	0.974
DemokritosGR2	0.956
BE-HM	*0.929*
ISI1	0.927
PolyU1	0.911
PolyU3	0.904
CLASSY2	0.901
CLASSY3	0.898
DemokritosGR1	0.897
ROUGE-2	*0.895*

Initial summaries (AllPeers)

Update summaries (AllPeers)

DemokritosGR2	0.968
Univille1	0.963
CLASSY1	0.960
CLASSY4	0.952
CLASSY2	0.946
BE-HM	0.944
NIRAJIIITH1	0.926
ISI1	0.868
ROUGE-2	*0.861*
PolyU1	0.821
PolyU3	0.814

Responsiveness

Initial summaries (AllPeers)

NIRAJIIITH1	0.977
Univille1	0.973
DemokritosGR2	0.944
BE-HM	0.938
ISI1	0.928
ROUGE-2	0.918
CLASSY2	0.918
CLASSY4	0.916
CLASSY1	0.914
ROUGE-SU4	0.909
DemokritosGR1	0.909

DemokritosGR2	0.977
Univille1	0.968
CLASSY1	0.955
CLASSY4	0.945
CLASSY2	0.941
NIRAJIIITH1	0.928
BE-HM	0.926
ROUGE-2	0.858
ISI1	0.850
PolyU1	0.797
DemokritosGR1	0.797

Update summaries (AllPeers)

NIST

Evaluation – Discriminative power

Initial summaries				Update summaries			
ID	difference (max 344)	no difference (max 0)	contradiction	ID	difference (max 344)	no difference (max 0)	contradiction
CIST1	344	0	0	CIST2	344	0	0
CIST2	344	0	0	DemokritosGR2	344	0	0
DemokritosGR2	344	0	0	NIRAJIIITH1	344	0	0
NIRAJIIITH1	344	0	0	Univille1	344	0	0
Univille1	344	0	0	CIST1	341	0	0
ROUGE-SU4	256	0	0	BE-HM	260	0	0
BE-HM	222	0	0	ROUGE-SU4	205	0	0
ROUGE-2	219	0	0	ROUGE-2	190	0	0

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

Initial summaries				Update summaries			
ID	difference (max 344)	no difference (max 0)	contradiction	ID	difference (max 344)	no difference (max 0)	contradiction
CIST1	344	0	0	CIST2	344	0	0
CIST2	344	0	0	DemokritosGR2	344	0	0
DemokritosGR2	344	0	0	NIRAJIIITH1	344	0	0
NIRAJIIITH1	344	0	0	Univille1	344	0	0
Univille1	344	0	0	CIST1	341	0	0
ROUGE-SU4	256	0	0	BE-HM	260	0	0
BE-HM	222	0	0	ROUGE-SU4	205	0	0
ROUGE-2	219	0	0	ROUGE-2	190	0	0

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

NIST

Evaluation – Discriminative power

Initial summaries				Update summaries			
ID	difference (max 318)	no difference (max 585)	contradiction	ID	difference (max 227)	no difference (max 676)	contradiction
ICL_SUM2	312	484	0	ICL_SUM2	218	501	0
PolyU3	303	489	0	PolyU4	215	524	0
DemokritosGR2	300	529	0	NIRAJIIITH1	212	559	0
ROUGE-SU4	299	532	0	DemokritosGR1	212	530	0
DemokritosGR1	298	515	0	PolyU2	209	537	0
CLASSY2	295	502	0	ROUGE-2	198	580	0
ROUGE-2	278	561	0	ROUGE-SU4	197	586	0
BE-HM	224	573	0	BE-HM	177	639	0

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

Initial summaries				Update summaries			
ID	difference (max 294)	no difference (max 609)	contradiction	ID	difference (max 208)	no difference (max 695)	contradiction
ICL_SUM2	287	483	0	ICL_SUM2	200	502	0
PolyU3	280	491	0	PolyU4	197	525	0
ROUGE-SU4	279	536	0	DemokritosGR1	193	530	0
CLASSY2	278	509	0	CLASSY4	192	579	0
DemokritosGR2	277	530	0	PolyU3	192	514	0
CLASSY3	276	545	0	ROUGE-2	185	586	0
ROUGE-2	259	566	0	ROUGE-SU4	181	589	0
BE-HM	212	585	0	BE-HM	163	644	0

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

NIST

AESOP task - Conclusions

- Correlations with Pyramid and Responsiveness
  - high correlations for baselines: ROUGE-2, ROUGE-SU4, BE-HM
  - high correlations for candidate metrics: CLASSY, DemokritosGR
  - better for NoModels than for AllPeers
- Similarity of discriminative power
  - limited for baselines: ROUGE-2, ROUGE-SU4, BE-HM
  - high for metrics: CIST, DemokritosGR, Univille, NIRAJIITH, PolyU, ICL_SUM, CLASSY
- Next year:
  - try to raise AllPeers correlations?
  - metric to emulate Readability?

NIST

Thank you