Overview of TAC 2010 Summarization Track

Guided 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

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

D1004A - Papua Tsunami
D1012C - Helios Crash
D1017D - Hurricane Floyd
D1023E - Austrian Avalanches
D1032F - Offshore Gas Leak

7 topics

2. Attacks (Criminal/Terrorist)

D1001A - Columbine Massacre
D1019D - Malaysia Resort Abduction
D1036G - Dioxin Poisoning Yushchenko
D1024E - Bomb Khartoum
D1029F - Baluchistan Uprising

7 topics

3. Health and Safety

D1005A - Parkinson's Disease
D1044H - Red Food Dye
D1006A - Vioxx
D1011C - Eating Disorders

12 topics

4. Endangered Resources

D1003A - Giant Panda
D1008B - Chesapeake Bay
D1015C - Rainforest Destruction
D1041H - Coral Reefs

10 topics

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

D1002A - Diallo Trial
D1016C - Soeharto Investigation
D1033C - South Korean Wire Tapping
D1040G - Robert Blake Murder Trial
D1042H - Lynndie England

10 topics

NIST

Guided Summarization task

- 8 NIST assessors
- 46 topics
- 20 documents selected for each topic
  - AQUAINT: New York Times, Associated Press, Xinhua News Agency (1999-2000; 1996-2000 for Xinhua documents)
  - AQUAINT-2: Agence France Presse, Central News Agency (Taiwan), Xinhua News Agency, Los Angeles Times-Washington Post News Service, New York Times, Associated Press (Oct 2004 - Mar 2006)
- 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:
  - 23 teams
  - 41 runs (up to two runs per team)
- 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

NIST

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)

  1. Extract all “information nuggets” (Summary Content Units/SCUs) from model summaries

D1016C-A
SCU: Soeharto is the former president of Indonesia
contr1: Soeharto, former president of Indonesia
contr2: Indonesian...former President Soeharto
contr3: Indonesia's former president Soeharto
contr4: Indonesian former President Soeharto's

D1016C-A
SCU: Soeharto thought to be extremely wealthy
contr1: Estimates of Soeharto's wealth range from $20 to $40 billion
contr2: The Indonesian Data Center reported Soeharto family assets of about U.S. $17.5 billion
contr3: Soeharto's wealth, once estimated to be billions of U.S. dollars
contr4: amassed a fortune worth billions of US dollars

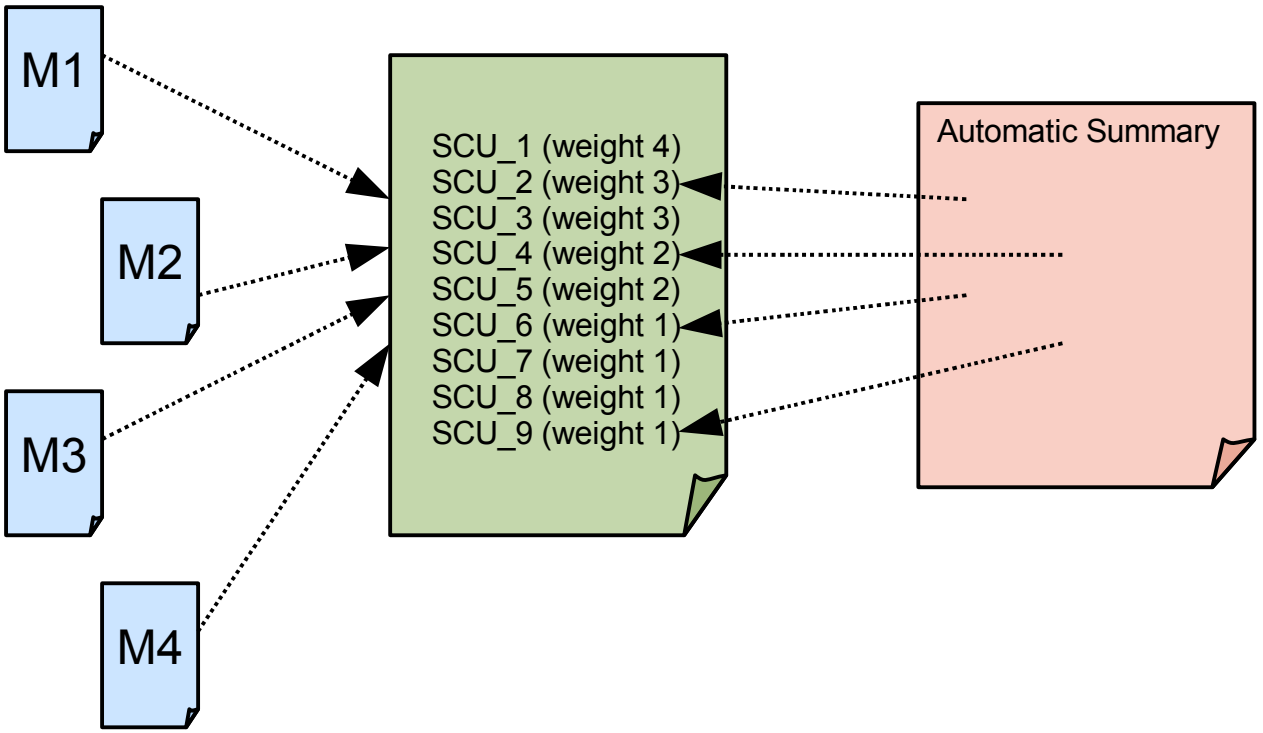
  2. SCU’s weight = number of model summaries that contain it

Guided Summarization task - Evaluation

- Pyramid (Passonneau et al., 2005)

  3. Check how many SCUs are present in the candidate summary

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



$$\frac{3 + 2 + 1 + 1}{4+3+3+2+2+1} = 0.467$$

Evaluation - Responsiveness

ID	Score	
E	4.9565	A
F	4.9130	A
G	4.8261	A
A	4.7826	A
H	4.6957	A
B	4.6957	A
D	4.6522	A
C	4.5652	A
CLASSY1	3.1739	B
Siel_101	3.1304	B C
HZNU2	3.0870	B C D
Guelph_NLP1	3.0652	B C D E
THU_HUANG2	3.0217	B C D E F
Guelph_NLP2	3.0217	B C D E F
ICL_SUM1	3.0217	B C D E F
CLASSY2	3.0217	B C D E F
PKUTM1	3.0217	B C D E F
ICL_SUM2	2.9783	B C D E F
JRC1	2.9783	B C D E F
THU_HUANG1	2.9783	B C D E F

models

ID	Score	
C	4.8261	A
B	4.7826	A
A	4.7826	A
F	4.7391	A
E	4.7391	A
G	4.6522	A
D	4.6522	A
H	4.5217	A
CLASSY1	2.7174	B
CLASSY2	2.6522	B C
ICTCAS2	2.5435	B C D
ICTCAS1	2.5217	B C D E
HZNU1	2.5217	B C D E
ICL_SUM2	2.4783	B C D E F
Baseline2	2.4783	B C D E F
THU_HUANG2	2.4565	B C D E F G
Siel_101	2.4565	B C D E F G
Guelph_NLP1	2.3913	B C D E F G H
ICL_SUM1	2.3913	B C D E F G H
seme12	2.3696	B C D E F G H

Initial summaries

Update summaries

Evaluation - Readability

ID	Score	
E	5.0000	A
D	5.0000	A
H	4.9565	A
F	4.9565	A
A	4.9130	A
B	4.8696	A
C	4.8261	A
G	4.7391	A
Baseline1	3.6522	B
CLASSY1	3.4565	B C
CLASSY2	3.3696	B C D
JRC1	3.3478	B C D
THU_HUANG2	3.3043	B C D E
JRC2	3.2826	B C D E
PKUTM1	3.2826	B C D E
Guelph_NLP2	3.2391	B C D E F
Guelph_NLP1	3.2174	B C D E F G
seme12	3.2174	B C D E F G
ICTCAS1	3.1957	B C D E F G
THU_HUANG1	3.1522	B C D E F G

models

ID	Score	
D	4.9565	A
A	4.9130	A
C	4.8696	A
B	4.8261	A
H	4.7826	A
F	4.7826	A
E	4.7826	A
G	4.6522	A
Baseline1	3.7391	B
CLASSY1	3.3261	B C
JRC2	3.3043	B C
Guelph_NLP1	3.2826	B C
THU_HUANG1	3.2826	B C
CLASSY2	3.2391	B C D
THU_HUANG2	3.1304	B C D E
seme11	3.1304	B C D E
seme12	3.1304	B C D E
Guelph_NLP2	3.1087	B C D E
JRC1	3.0870	B C D E F
Siel_101	3.0217	C D E F

Initial summaries

Update summaries

Evaluation - Pyramid

ID	Score		ID	Score	
F	0.88813	A	F	0.75957	A
G	0.87170	A	G	0.74813	A
H	0.80839	A B	B	0.72922	A
E	0.80761	A B	H	0.67357	A B
A	0.77917	A B	D	0.66909	A B
B	0.74665	A B	A	0.62865	A B
D	0.72043	A B	E	0.62635	A B
C	0.66109	B	C	0.55135	B
Siel_101	0.41754	C	CLASSY1	0.31587	C
CLASSY2	0.40709	C D	CLASSY2	0.31389	C D
Guelph_NLP1	0.40611	C D	Siel_101	0.27487	C D E
dataminer1	0.39272	C D E	ICL_SUM2	0.27413	C D E
CLASSY1	0.39241	C D E	HZNU1	0.27091	C D E
ICL_SUM1	0.39172	C D E	ICTCAS1	0.26943	C D E
THU_HUANG2	0.39059	C D E	THU_HUANG2	0.26915	C D E
HZNU2	0.38652	C D E	ICTCAS2	0.26739	C D E F
seme12	0.38217	C D E F	Baseline2	0.25576	C D E F G
dataminer2	0.37948	C D E F	ICL_SUM1	0.25572	C D E F G
HZNU1	0.37880	C D E F	Guelph_NLP2	0.25174	C D E F G H
PKUTM1	0.37839	C D E F	JRC2	0.24322	C D E F G H

models

Initial summaries

Update summaries

Evaluation – Average scores

Pyramid	Models	Automatic
Initial summaries	0.785	0.302
Update summaries	0.673	0.199

Responsiveness	Models	Automatic
Initial summaries	4.761	2.565
Update summaries	4.712	2.102

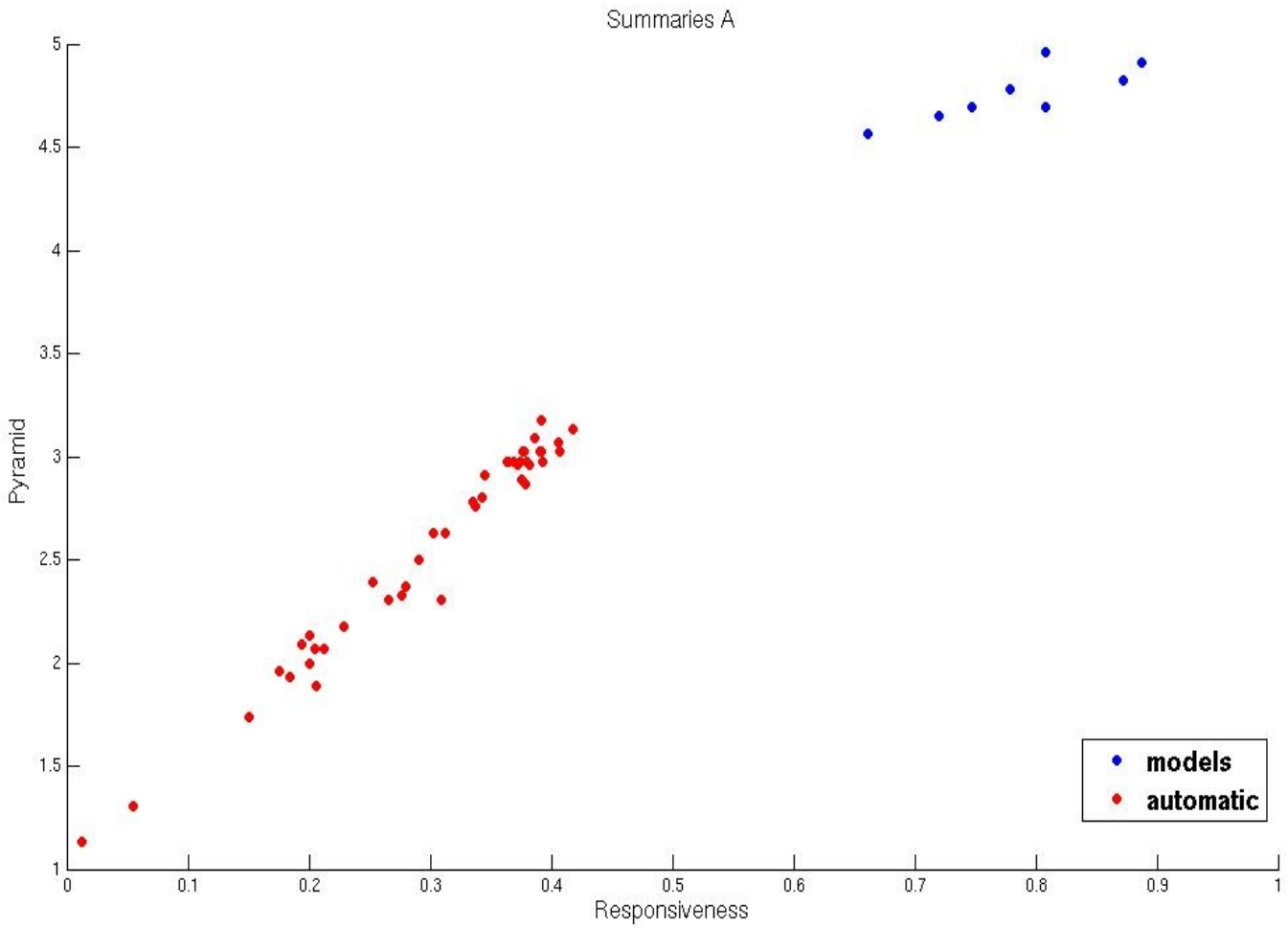
no significant difference

SCUs	Models	Automatic
Initial summaries	11.663*	4.385*
Update summaries	7.331*	2.281*

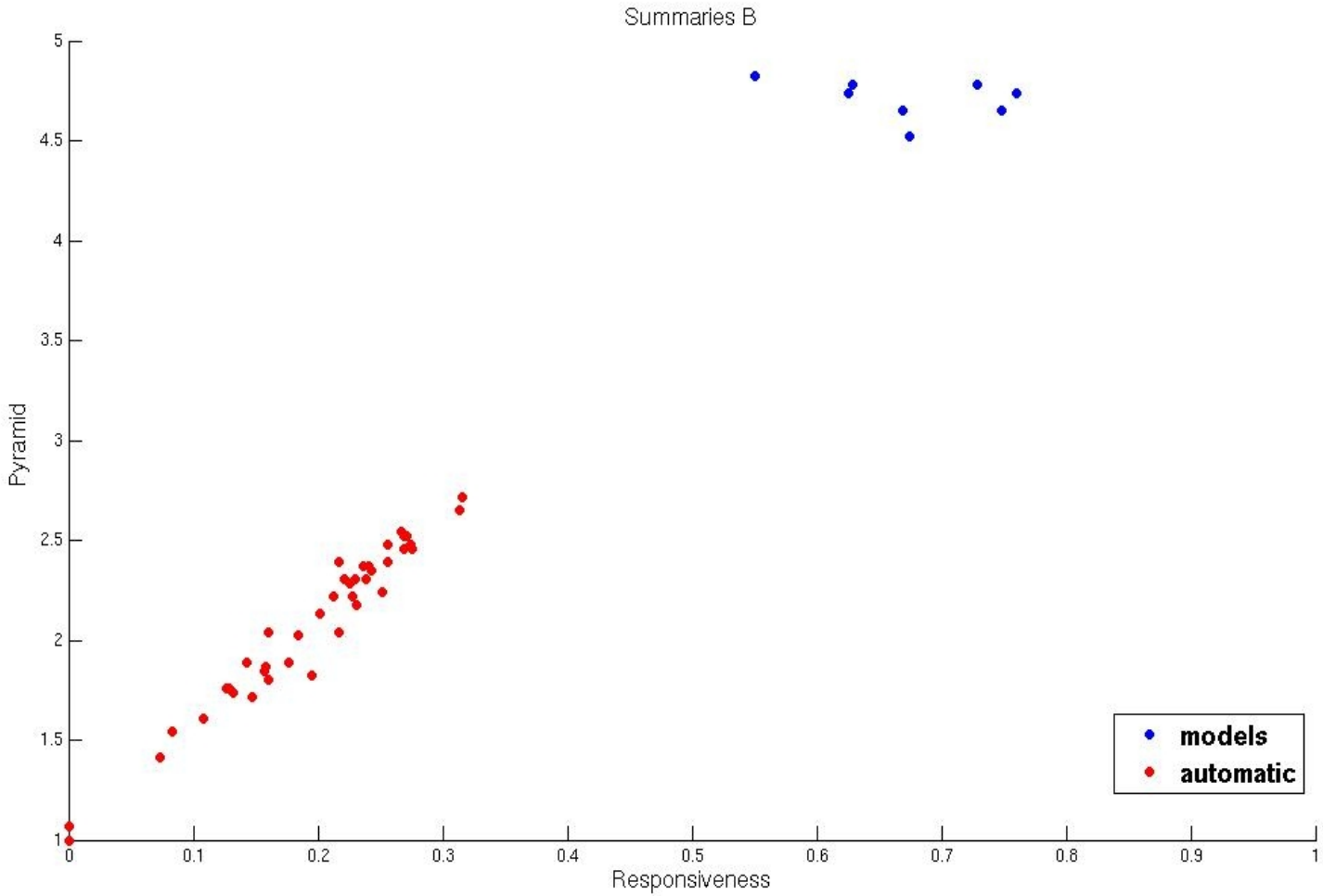
*p > 0.05

Readability	Models	Automatic
Initial summaries	4.908	2.837
Update summaries	4.821	2.765

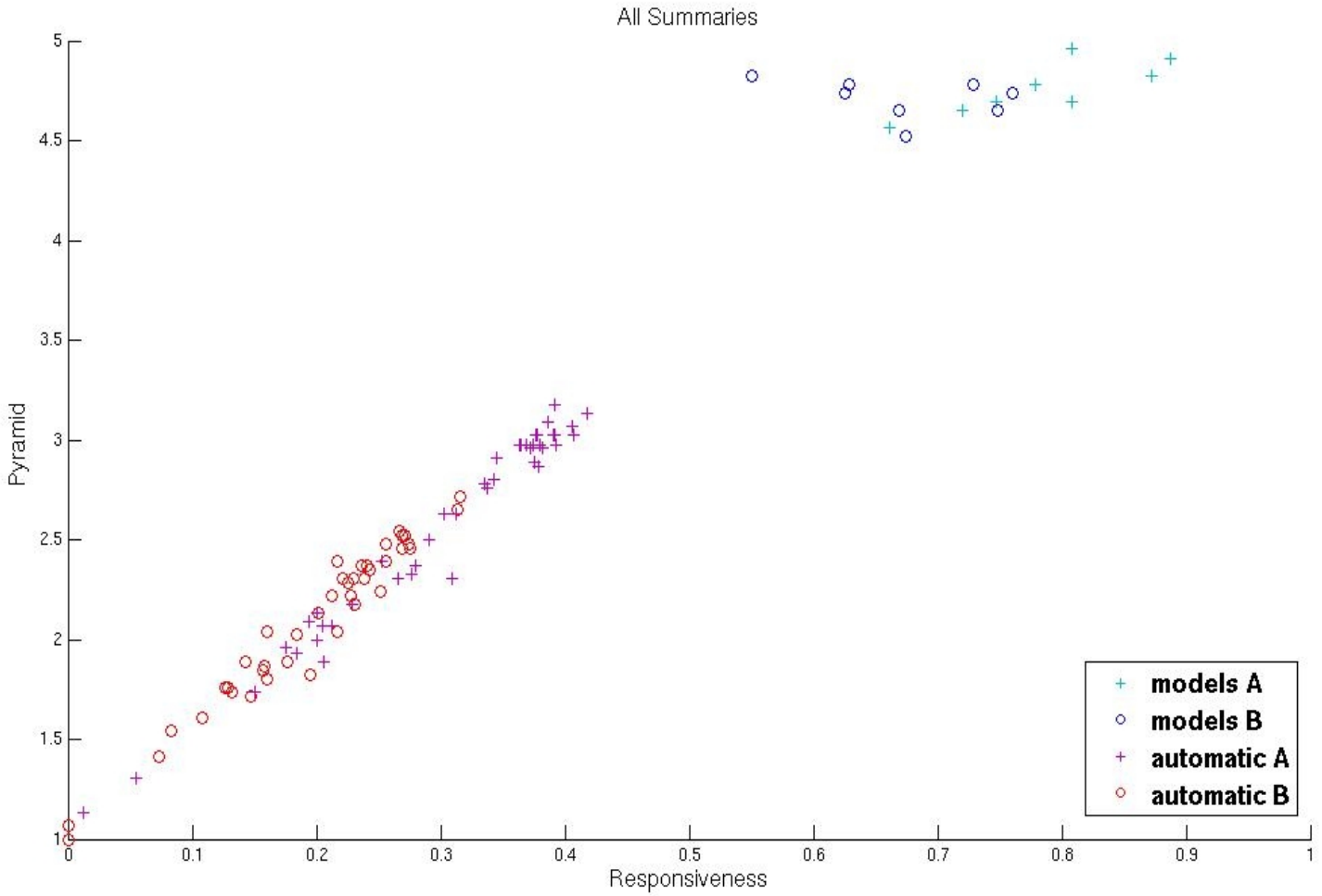
Evaluation – Averages (Initial summaries)



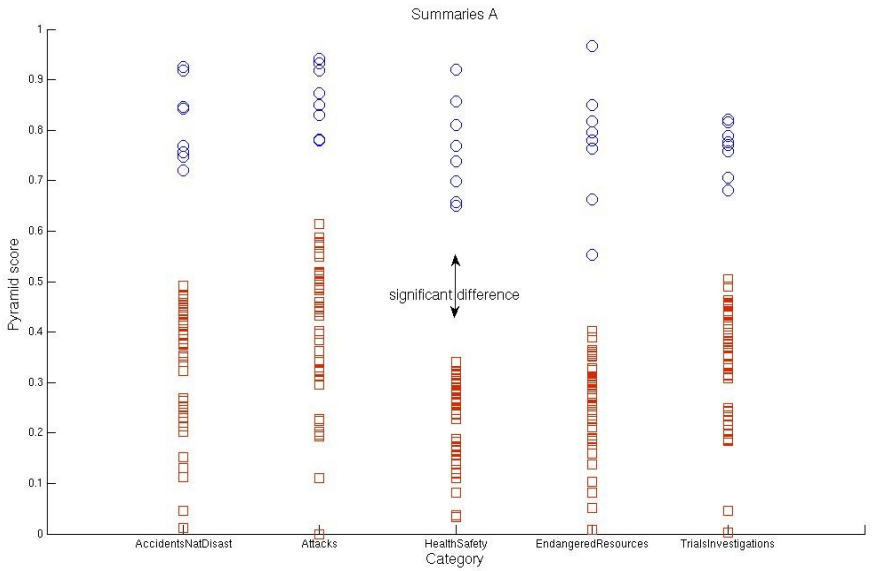
Evaluation – Averages (Update summaries)



Evaluation – Averages (All summaries)

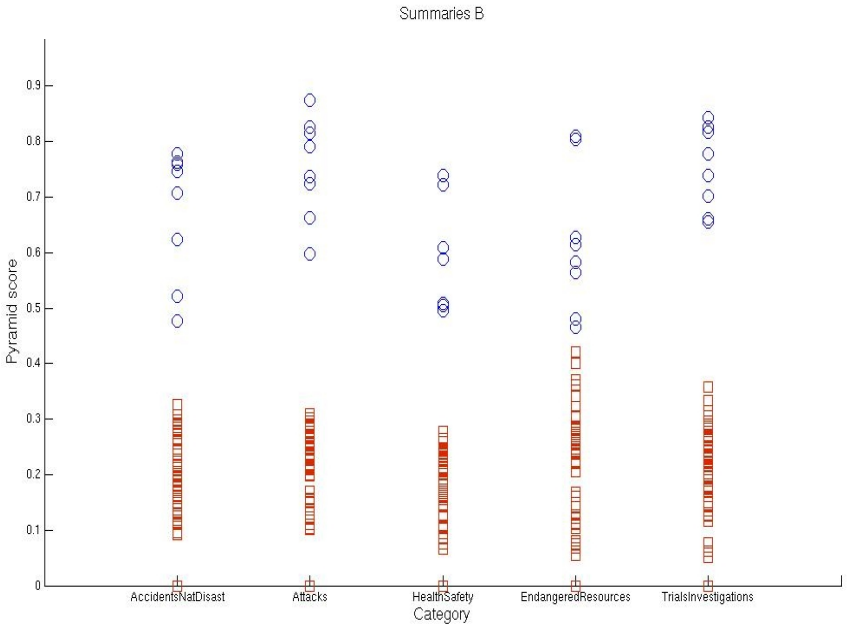


Evaluation – Categories – Pyramid



Initial Summaries

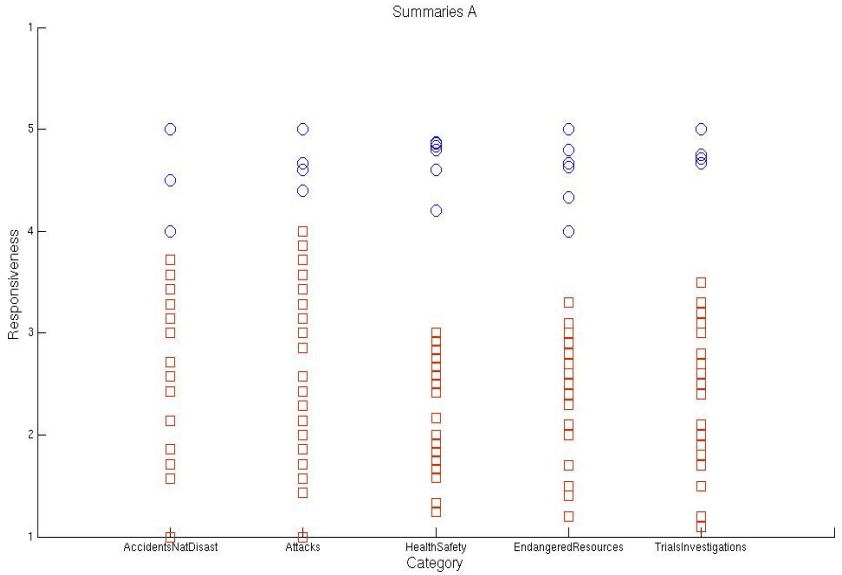
Cat	Human		Cat	Automatic	
Attacks	0.857	*A*	Attacks	0.410	*A*
Acc&NatDis	0.812	*AB*	Trials&Inv	0.342	*B*
EndangRes	0.773	*AB*	Acc&NatDis	0.334	*B*
Health&Safety	0.767	*AB*	EndangRes	0.256	*C*
Trials&Inv	0.751	*B*	Health&Safety	0.224	*D*



Update summaries

Cat	Human		Cat	Automatic	
Trials&Inv	0.749	*A*	EndangRes	0.216	*A*
Attacks	0.745	*AB*	Attacks	0.208	*AB*
Acc&NatDis	0.700	*AB*	Trials&Inv	0.206	*ABC*
Health&Safety	0.610	*C*	Acc&NatDis	0.192	*ABCD*
EndangRes	0.604	*C*	Health&Safety	0.177	*D*

Evaluation – Categories – Responsiveness

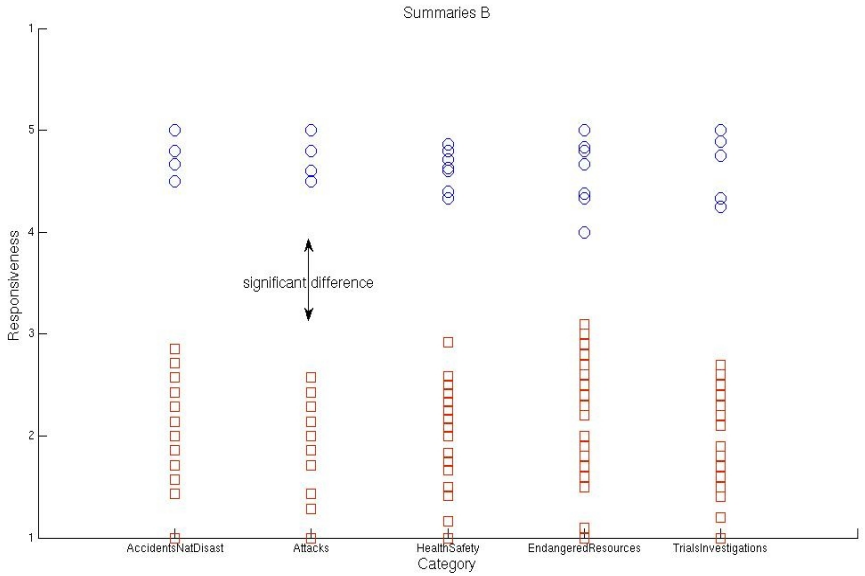


Initial summaries

Cat	Human		Cat	Automatic	
Trials&Inv	4.825	A	Acc&NatDis	2.877	A
Acc&NatDis	4.821	AB	Attacks	2.797	AB
Attacks	4.786	ABC	Trials&Inv	2.593	BC
Health&Safety	4.750	ABCD	EndangRes	2.505	C
EndangRes	4.650	ABCD	Health&Safety	2.275	D

Update summaries

Cat	Human		Cat	Automatic	
Attacks	4.857	A	EndangRes	2.179	A
Trials&Inv	4.825	AB	Health&Safety	2.137	AB
Acc&NatDis	4.714	ABC	Acc&NatDis	2.066	ABC
Health&Safety	4.625	ABCD	Trials&Inv	2.046	ABCD
EndangRes	4.600	ABCD	Attacks	2.043	ABCD



Guided Summarization task - Conclusions

- Quality gap between human and automatic summarizers
  - smaller gap for some categories than others
    - Pyramid (i.e., content similarity to models): Endangered Resources (human summarizers low in similarity)
    - Responsiveness (i.e., relevance to topic): Attacks, Accidents and Natural Disasters

- Use of categories and aspects in summarization
  - finding aspect-related words/sentences from provided samples or training corpora
  - NER for WHO, WHERE, WHEN aspects
  - IE/event extraction systems for some aspects
  - categories/aspects as query expansion terms

NIST

Thank you