An Assessment of the Accuracy of Automatic Evaluation in Summarization

Karolina Owczarzak, NIST
John M. Conroy, IDA Center for Computing Sciences
Hoa Trang Dang, NIST
Ani Nenkova, University of Pennsylvania

Just How Good is that Summary?

- Manual Metrics
  - Readability: Qualitative score of linguistic quality.
  - Responsiveness: Qualitative score of overall responsiveness to the given task.
  - Pyramid: A quantitative measure of content.

- Automatic Metrics
  - ROUGE-1,2,SU4, (with & w/o stop word removal)
  - AESOP 2011, BEwT-E

Challenges

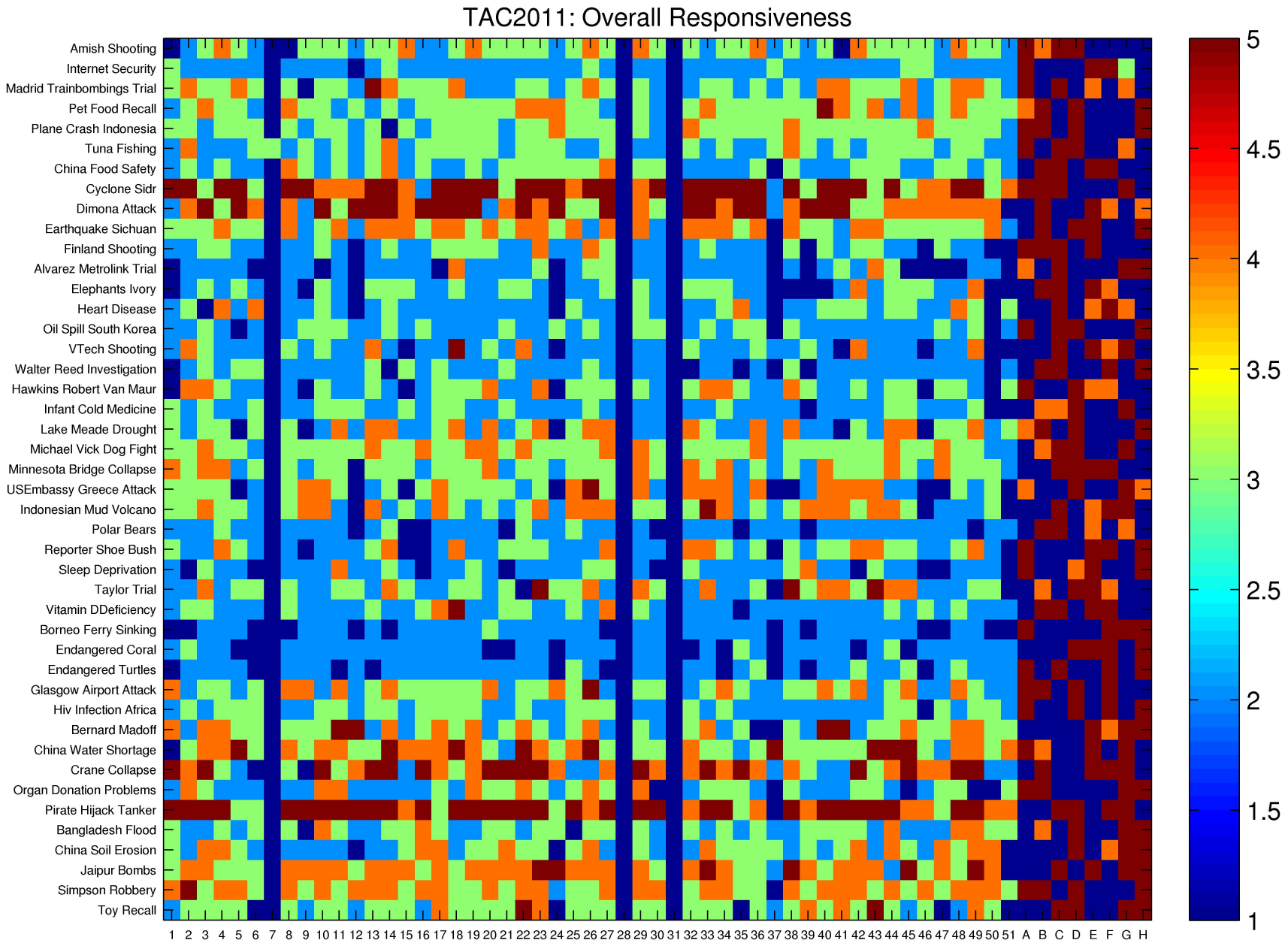
- Summarization systems are evaluated by evaluating each summary on a topic.
- However,
  - Topics differ in difficulty to summarizer.*
  - Humans judge inconsistently.⌘
  - Human evaluation is expensive.
- Desire to rank summarization systems.
  - Traditionally, average scores are produced.

* Nenkova & Louis, Can You Summarize This?, ACL 2008,

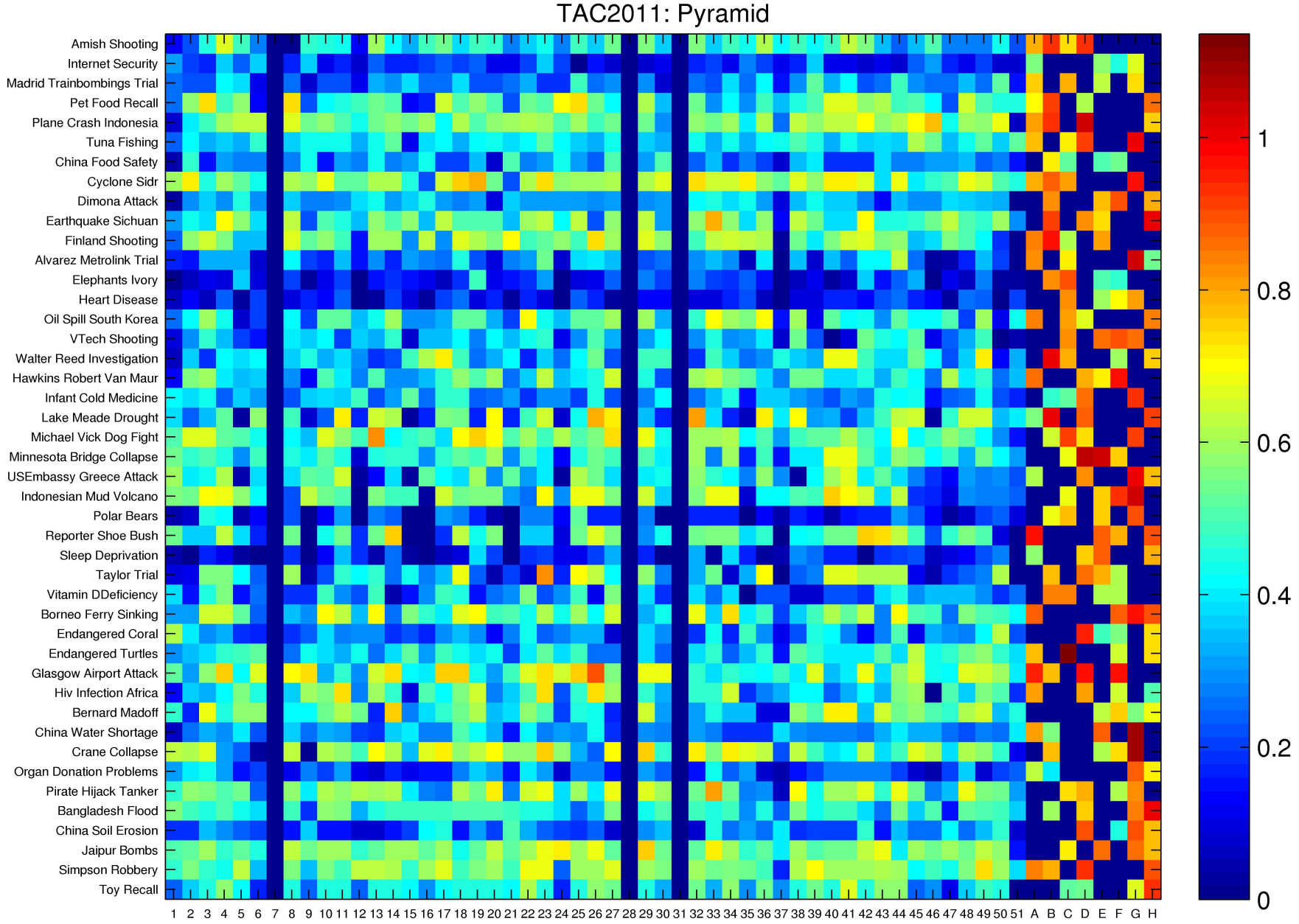
⌘ Owczarzak, Dang, Rankel & Conroy, Assessing the Effect of Inconsistent Assessors on Summarization Evaluation, ACL 2012.

What Makes a Automatic Good Metric?

- Past:
  - Correlation measures, e.g. Pearson, Spearman, Kendall Tau.
- Proposal:
  - Estimate the probability that an automatic metric will *agree* with a manual metric when comparing two *systems* when taking *statistical significance* into account.



Thanks to Peter Rankel for this slide and the next too!


TAC2011: Pyramid
Amish Shooting
Internet Security
Madrid Trainbombings Trial
Pet Food Recall
Plane Crash Indonesia
Tuna Fishing
China Food Safety
Cyclone Sidr
Dimona Attack
Earthquake Sichuan
Finland Shooting
Alvarez Metrolink Trial
Elephants Ivory
Heart Disease
Oil Spill South Korea
VTech Shooting
Walter Reed Investigation
Hawkins Robert Van Maur
Infant Cold Medicine
Lake Meade Drought
Michael Vick Dog Fight
Minnesota Bridge Collapse
USEmbassy Greece Attack
Indonesian Mud Volcano
Polar Bears
Reporter Shoe Bush
Sleep Deprivation
Taylor Trial
Vitamin DDeficiency
Borneo Ferry Sinking
Endangered Coral
Endangered Turtles
Glasgow Airport Attack
Hiv Infection Africa
Bernard Madoff
China Water Shortage
Crane Collapse
Organ Donation Problems
Pirate Hijack Tanker
Bangladesh Flood
China Soil Erosion
Jaipur Bombs
Simpson Robbery
Toy Recall
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 A B C D E F G H
1
0.8
0.6
0.4
0.2
0

How to compare systems?

- Simple *t*-test would wash out the variation in difficulty of comparing two summarization systems.
- Well known problem: Variation across data.
- Well known remedy: Paired testing, e.g. paired *t*-test (Mann-Whitney) or non-parametric Wilcoxon test.
- Rankel, Conroy, Slud, O'Leary EMNLP 2010 show paired testing gives many summarization metrics more *power*.

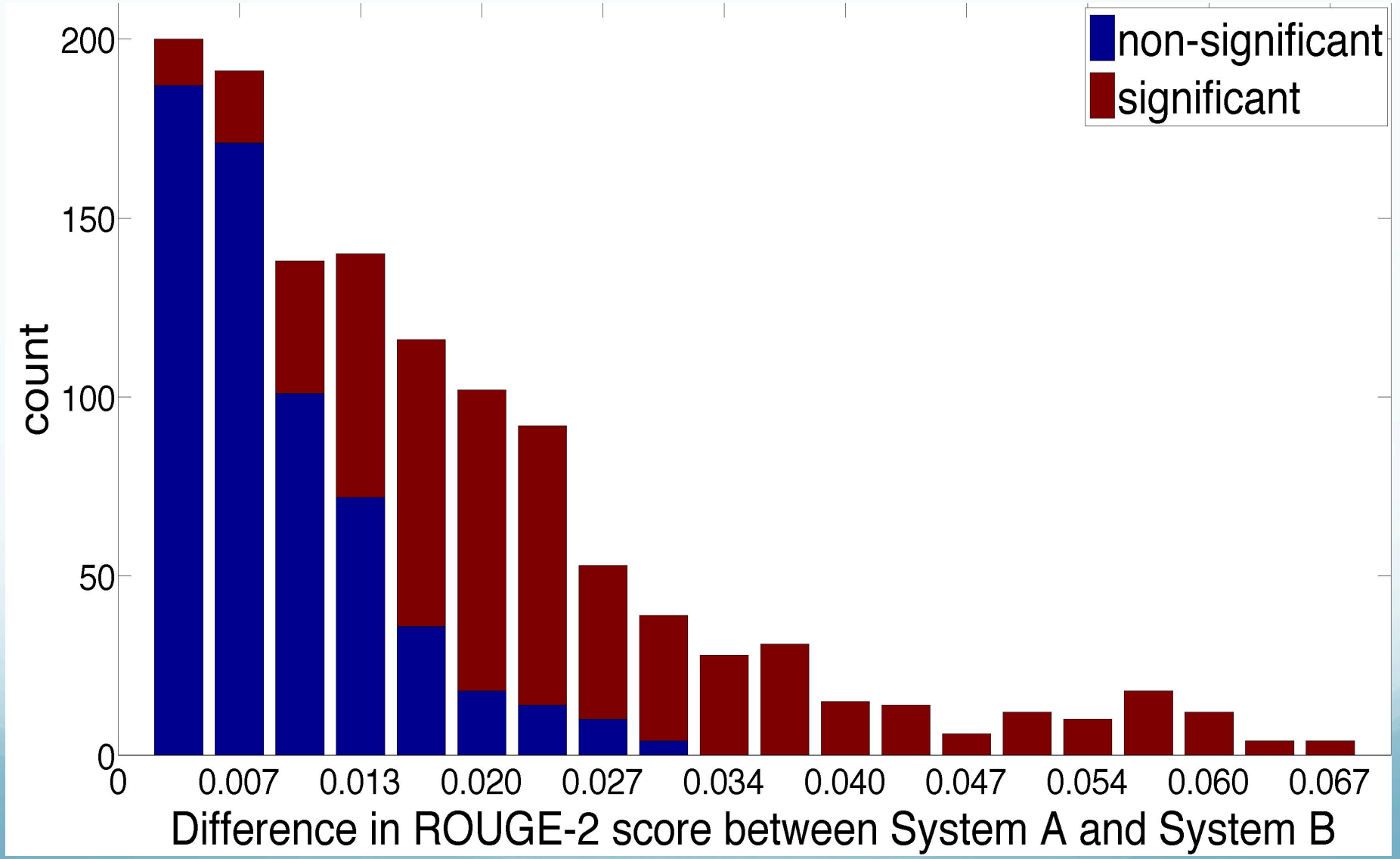
Our Hypothesis Test

H_0: median X-Y=0, X and Y are random variables corresponding to scores for two systems A and B. (A and B perform about the same.)

H_a: *median X-$Y \neq 0$.*
(A and B are significantly different!)

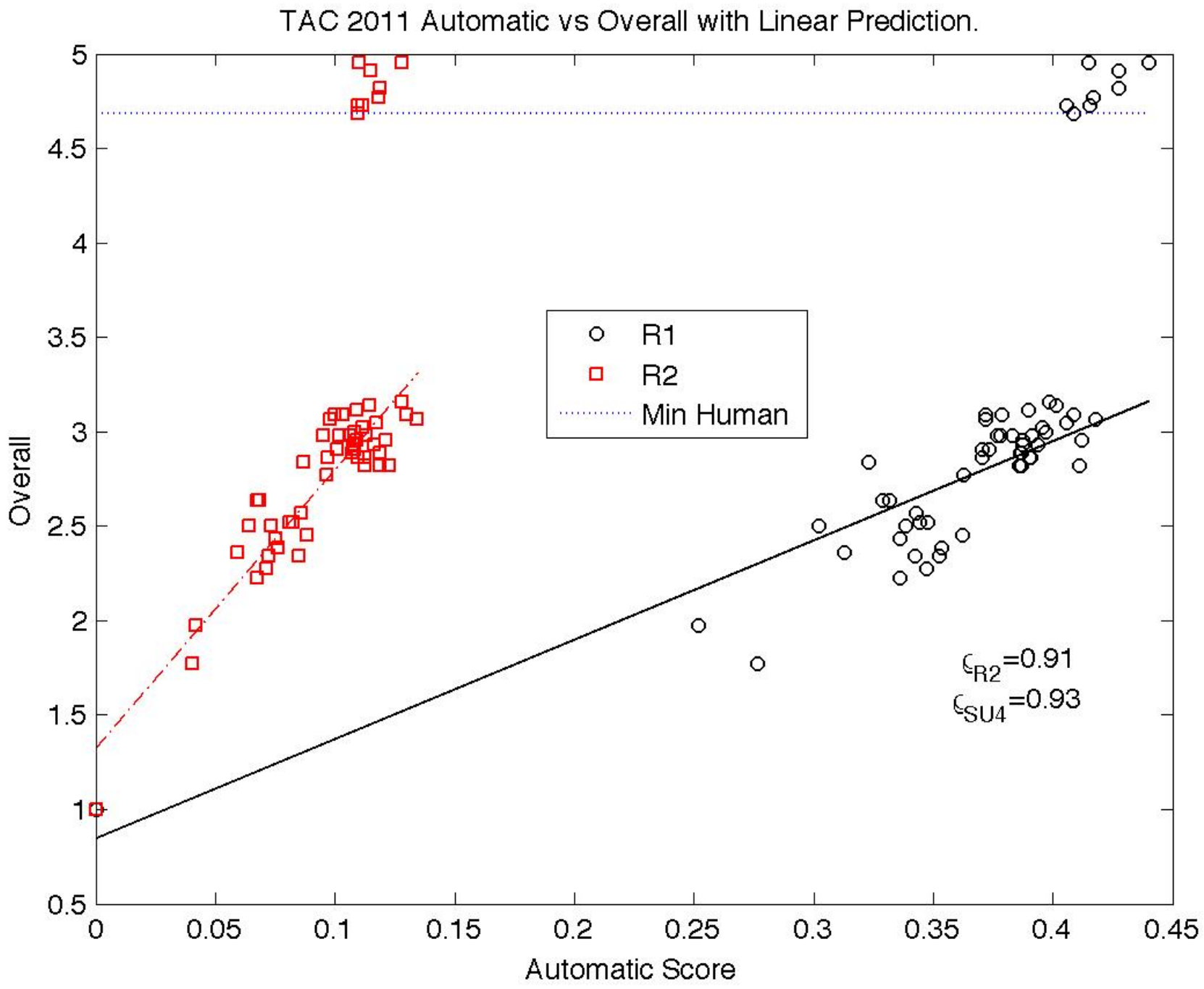
If median performance of A is greater than B and the null hypothesis is rejected, we say "*A significantly outperforms B.*"

How Much of of Difference Is Significant?



Comparing Metrics

Metric 1 Says	Metric 2 Says	Interpretation
$m(X\text{-}Y)=0$	$m(X\text{-}Y)=0$	Agree *X* and *Y* are about the same
$m(X\text{-}Y)\neq 0$	$m(X\text{-}Y)\neq 0$	Agree *X* and *Y* are different and $X>>Y$
$m(X\text{-}Y)\neq 0$	$m(X\text{-}Y)=0$	Disagreement
$m(X\text{-}Y)=0$	$m(X\text{-}Y)\neq 0$	Disagreement
$m(X\text{-}Y)=0$	$m(X\text{-}Y)=0$	metric 1 said $X>>Y$ & metric 2 $Y>>X$


TAC 2011 Automatic vs Overall with Linear Prediction.
Overall
Automatic Score
R1
R2
Min Human
ζR2=0.91
ζSU4=0.93
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45

Data

- Text Analysis Conference (TAC) 2008-2011

Year	Topics	Auto-Systems	Humans	Reference Summaries
2008	48	58	8	4
2009	44	55	8	4
2010	46	43	8	4
2011	44	50	8	4

Auto-metrics: ROUGE-1, 2, SU4 with and without stop word removal.
Manual Metrics: Pyramid and Overall Responsiveness

Metrics Performance for Comparing Auto-Systems

	Pyramid		Responsiveness	
	Sig	*All*	*Sig*	*All*
R1	0.77	0.87	0.70	0.82
R2	0.81	0.89	0.75	0.83
SU4	0.80	0.88	0.73	0.82

Sig: Pr(metric 1 agrees with metric 2 when they are significant difference between systems exists.

All: Pr(metric 1 agrees with metric 2 for both significant and non-significant differences between systems.

Metrics Performance on Comparing Auto vs. Humans

	Pyramid		Responsiveness	
	Sig	*All*	*Sig*	*All*
R1	0.90	0.99	0.90	0.99
R2	0.75	0.94	0.75	0.94
SU4	0.82	0.96	0.82	0.96

Sig: Pr(metric 1 agrees with metric 2 when they are significant difference between systems exists.

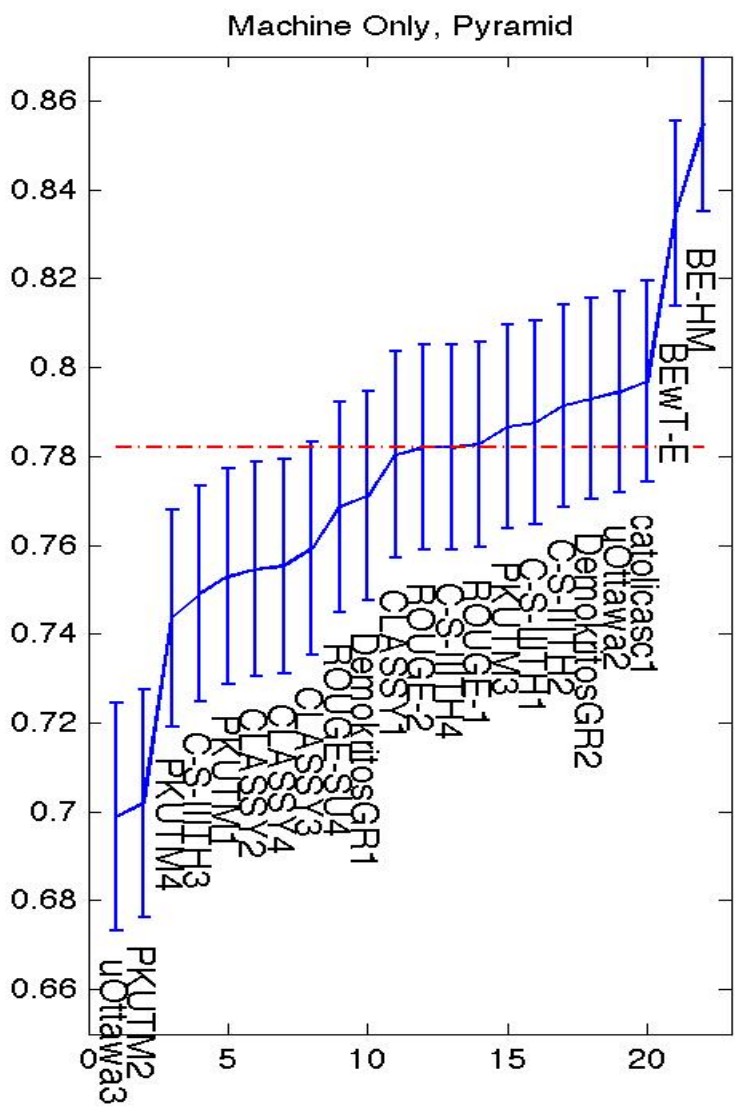
All: Pr(metric 1 agrees with metric 2 for both significant and non-significant differences between systems.

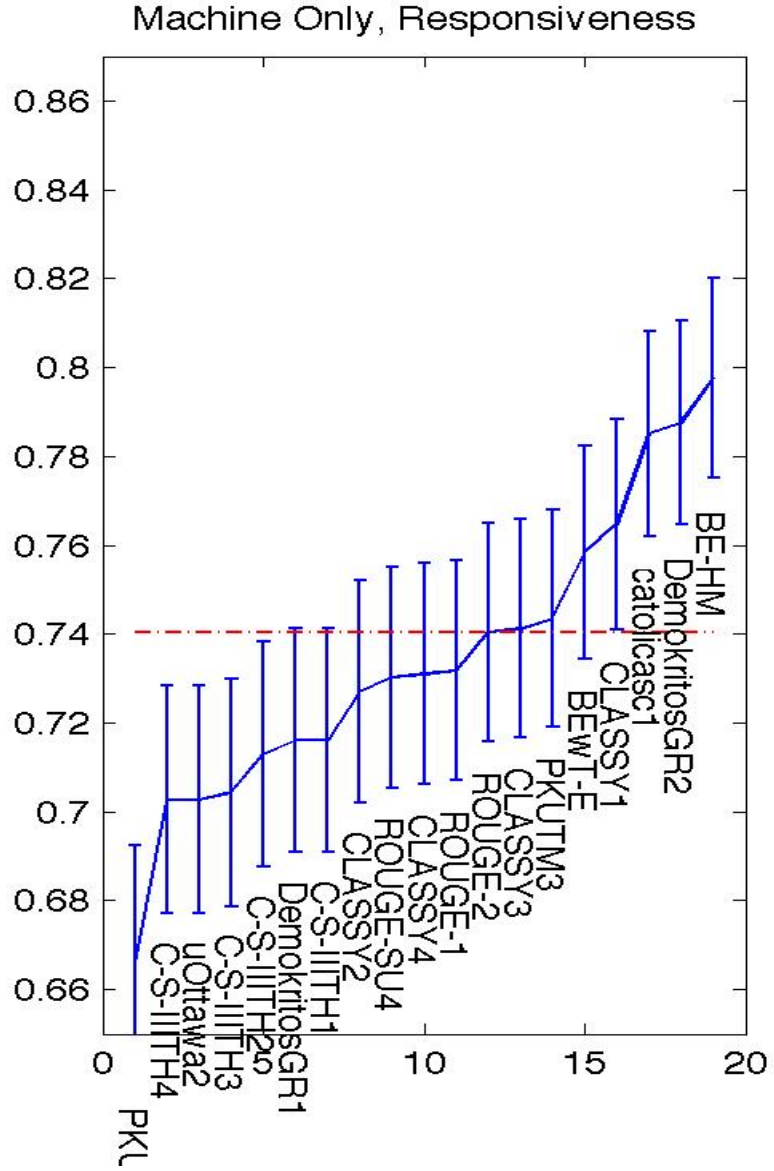
AESOP 2011

- Automatic Evaluation of Summaries of Peers, a metric "bakeoff."
- 25 official entries and ROUGE-1 and BEwT-E, (Basic Elements with Transformations for Evaluation)*
- Baselines:
  - ROUGE-2 for with automatic systems.
  - ROUGE-1 for between human vs automatic.

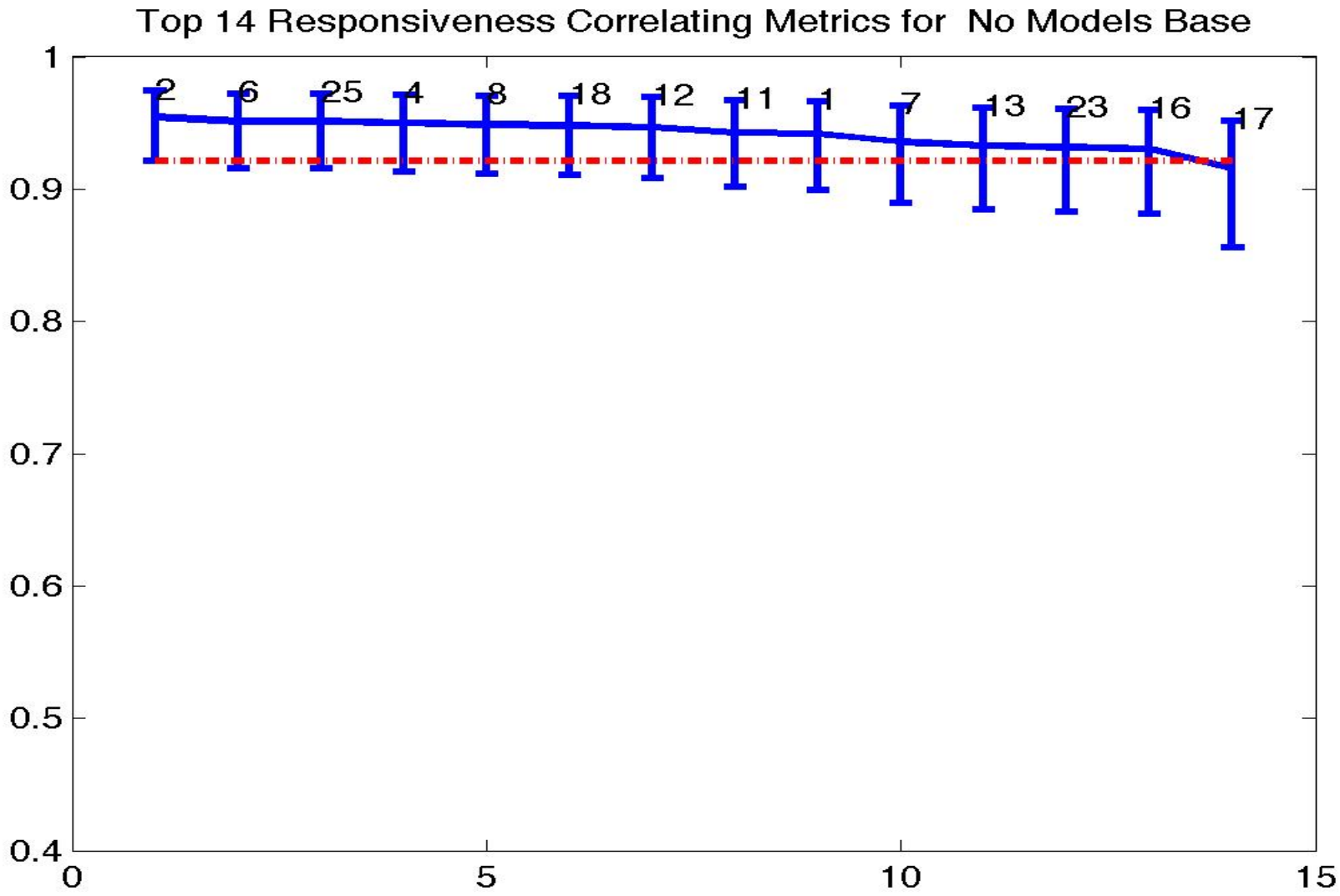
*Thanks to Stephen Tratz and Ed Hovy.

Comparing Automatic Summaries

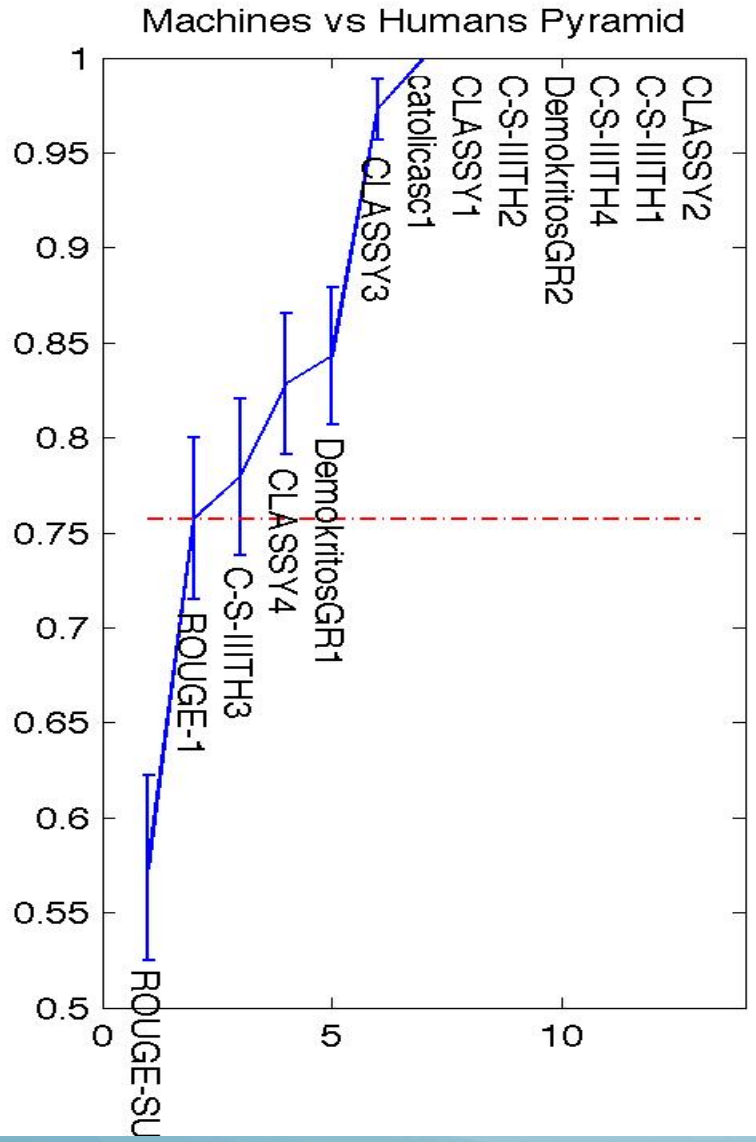


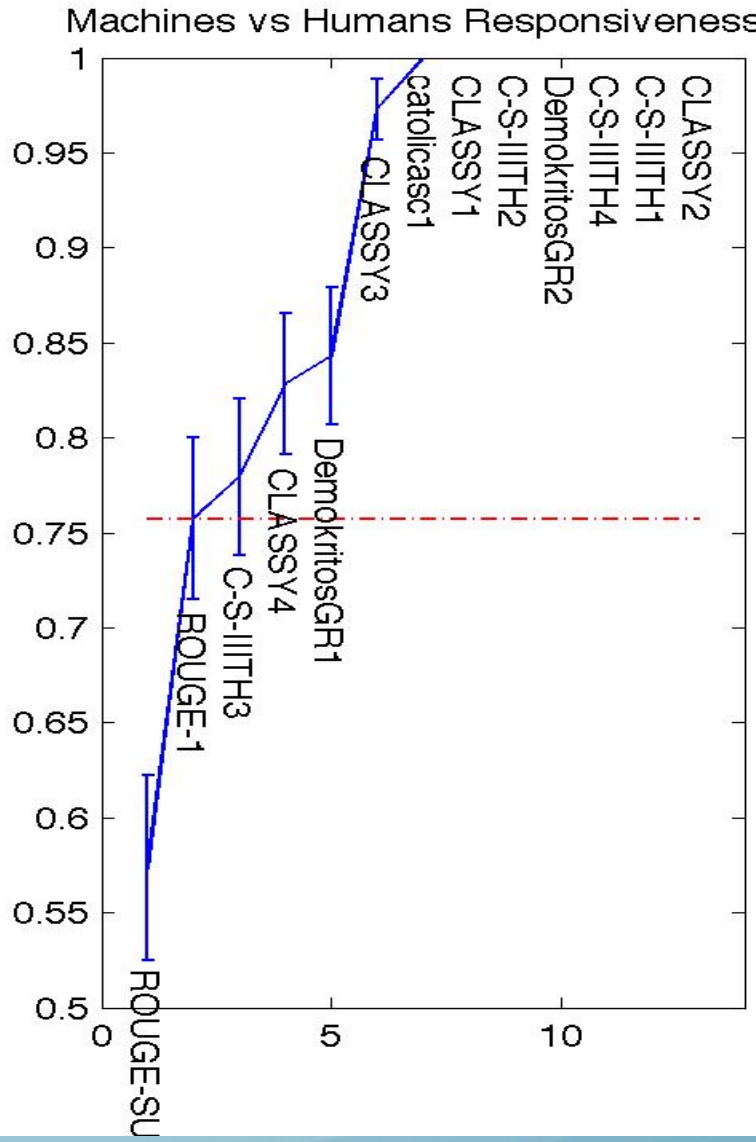


Ranking Based on Pearson



Comparing Automatic vs Humans





Summary

- Statistical significance is essential for comparing systems.
- Paired testing give more statistical power. Rankel, Conroy, Slud, O'Leary, EMNLP 2011.
- Is system *A significantly better than system B?*
- Evaluated an automatic metric by how well it agrees, taking significance into account with manual metric.