# CLASSY and TAC 2008 Metrics

*John M. Conroy*
Judith D. Schlesinger
IDA Center for Computing Sciences,USA

# Outline

- CLASSY 08
  - Update: System 6, 37, 60.
  - [Opinion: System 5, 36]
- What we submitted.
- How we did and how the metrics compare.
- Combining metrics.
- Meta-evaluation: evaluation of evaluation.

# CLASSY (Clustering, Linguistics, And Statistics for Summarization Yield)

- Linguistic preprocessing.
  - Shallow parsing
  - Find sentences and apply trimming techniques.
- Sentence Scoring.
  - Approximate Oracle.
- Redundancy Removal.
  - Select a subset of sentences.
  - LSI and non-negative “QR.”
- Ordering
  - TSP

# Linguistic Processing

- Eliminations
  - Gerund phrases
  - Relative clause appositives
  - Attributions
  - Lead adverbs and phrases
    - For example, On the other hand, …
  - Medial adverbs
    - too, however, …

# An Oracle Score

- An oracle might tell us Pr(*t*)
  Pr(*t*)=Probability that a human will choose term *t* to be included in a summary.
- If we had human summaries, we could estimate Pr(*t*) based on our data
  - E.g., 0, 1/4, 1/2, 3/4, or 1 if 4 human summaries are provided.
  - Oracle Score: fraction of expected abstract terms (vector space model).

# A Simple Approximation of $P(t|\tau)$

- We approximate $P(t|\tau)$ by

$$P_{sq\rho}(t \mid \tau) = \frac{1}{4}s(t) + \frac{1}{4}q(t) + \frac{1}{2}\rho(t)$$

$$s(t) = \begin{cases} 1 \text{ if } t \text{ is a signature term} \\ 0 \text{ if } t \text{ is not a signature term} \end{cases}$$

$$q(t) = \begin{cases} 1 \text{ if } t \text{ is a query term} \\ 0 \text{ if } t \text{ is not a query term} \end{cases}$$

$$\rho(t \mid \tau) = \text{ probability } t \text{ occurs in a sentence considered for selection.}$$

- The score of a sentence is the sum of $P(t|\tau)$ taken over its terms divided by its length.

# Smoothing and Redundancy Removal

Use approximate oracle to select candidate sentences (~3X *words*).

– Terms as sentence features
  - Terms: $\{t_1, \ldots, t_m\} \in \boldsymbol{R}^m$
  - Sentences: $\{s_1, \ldots, s_n\} \in \boldsymbol{R}^n$
  - Scaling: each column scaled to score.
  - LSI to reduce rank 0.65*n*.

– Non-negative "QR" to select sentences.

| | $s_1$ | $\cdots$ | $s_n$ |
|---|---|---|---|
| $t_1$ | $a_{11}$ | $\cdots$ | $a_{1n}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $t_m$ | $a_{m1}$ | $\cdots$ | $a_{mn}$ |

# Ordering Sentences

- Approximate TSP to increase flow.
- Start with worst...
- Order the lowest scoring sentence last.
- Order the other sentences so that the sum of the distances between adjacent sentences is minimized (TSP).
- $B_{ij}$ =number number words sentence *i* and *j* have in common.

$$c_{ij} = -\frac{b_{ij}}{\sqrt{b_{ii}}\sqrt{b_{jj}}}$$

# Adaptations for Update

- Sub-task A: run CLASSY on 10 docs.
- Sub-task B:
  - Use docs A and B to generate signature terms.
  - Project term-sentence matrix to orthogonal complement of submitted summary.
  - Select sentences from 10 new documents.
- This update strategy scored best  in 2007.

# Three Submissions

- System 6: background = AQUAINT 2 Complete Sentences: Bin packing to choose last sentence or two.
- System 37:background = AQUAINT 2 Possible Fragments
- System 60: background AQUAINT 1 Possible Fragments

# Content and Responsiveness

- DUC 2007 Main Task: Systems ending summary with sentence had significantly higher content responsiveness, Conroy & Dang 2008 COLING. However, content responsiveness "behaved like" overall responsiveness of 2006!
- DUC 2007 Update Task: Systems ending summary with sentence had significantly lower content responsiveness.

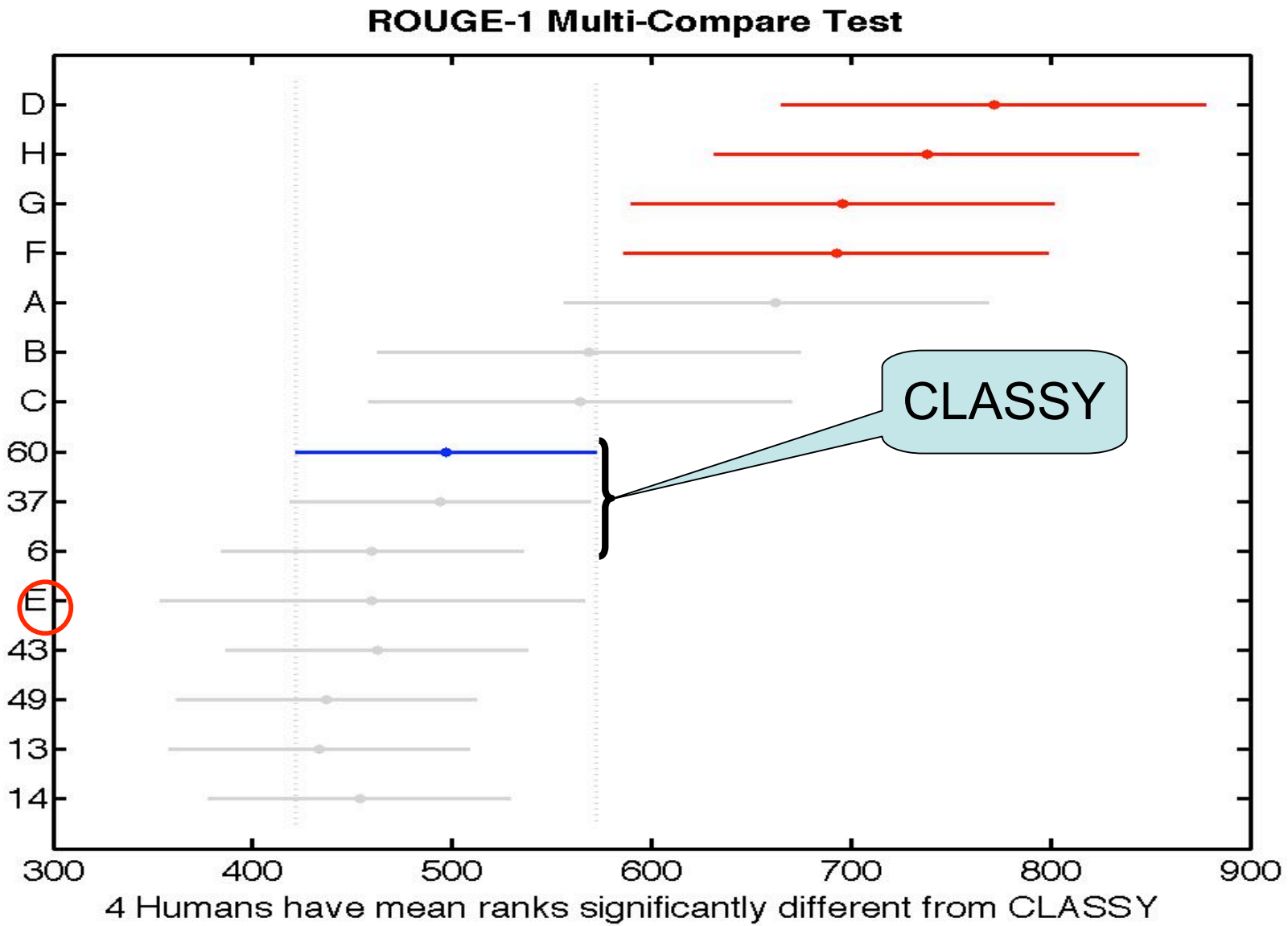
# 2008 Update Task

| Metric | Sentence | Fragment | *p*-Value |
|---|---|---|---|
| ROUGE-BE | 0.045 | 0.043 | 0.092 |
| ROUGE-2 | 0.073 | 0.072 | 0.319 |
| ROUGE-SU4 | 0.287 | 0.289 | 0.974 |
| Linguistic | 2.422 | 2.239 | 6.74e-8 |
| Pyramid | 0.232 | 0.233 | 0.838 |
| Over. Resp. | 2.203 | 2.137 | 0.010 |

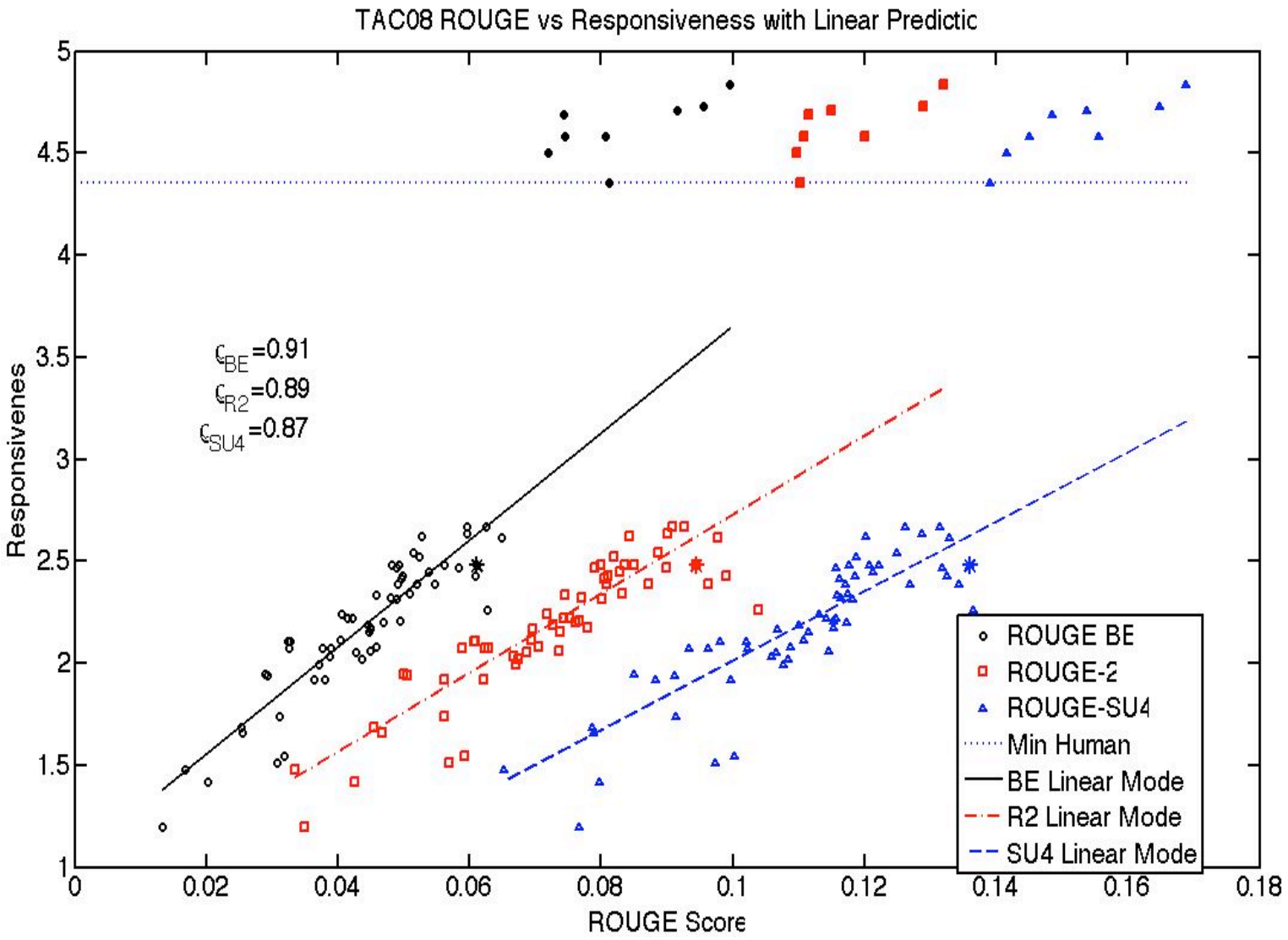
# What about CLASSY?

- CLASSY
  - Pyramid, Responsiveness, ROUGE-BE
    - No significant difference between submissions.
  - ROUGE 2, SU4
    - Ending with fragment significantly higher.
  - *No significant difference background model: AQUAINT 1 vs. 2.*
- Conclusions:
  - Perhaps we could do better bin packing!
  - *Signature terms are relatively robust.*

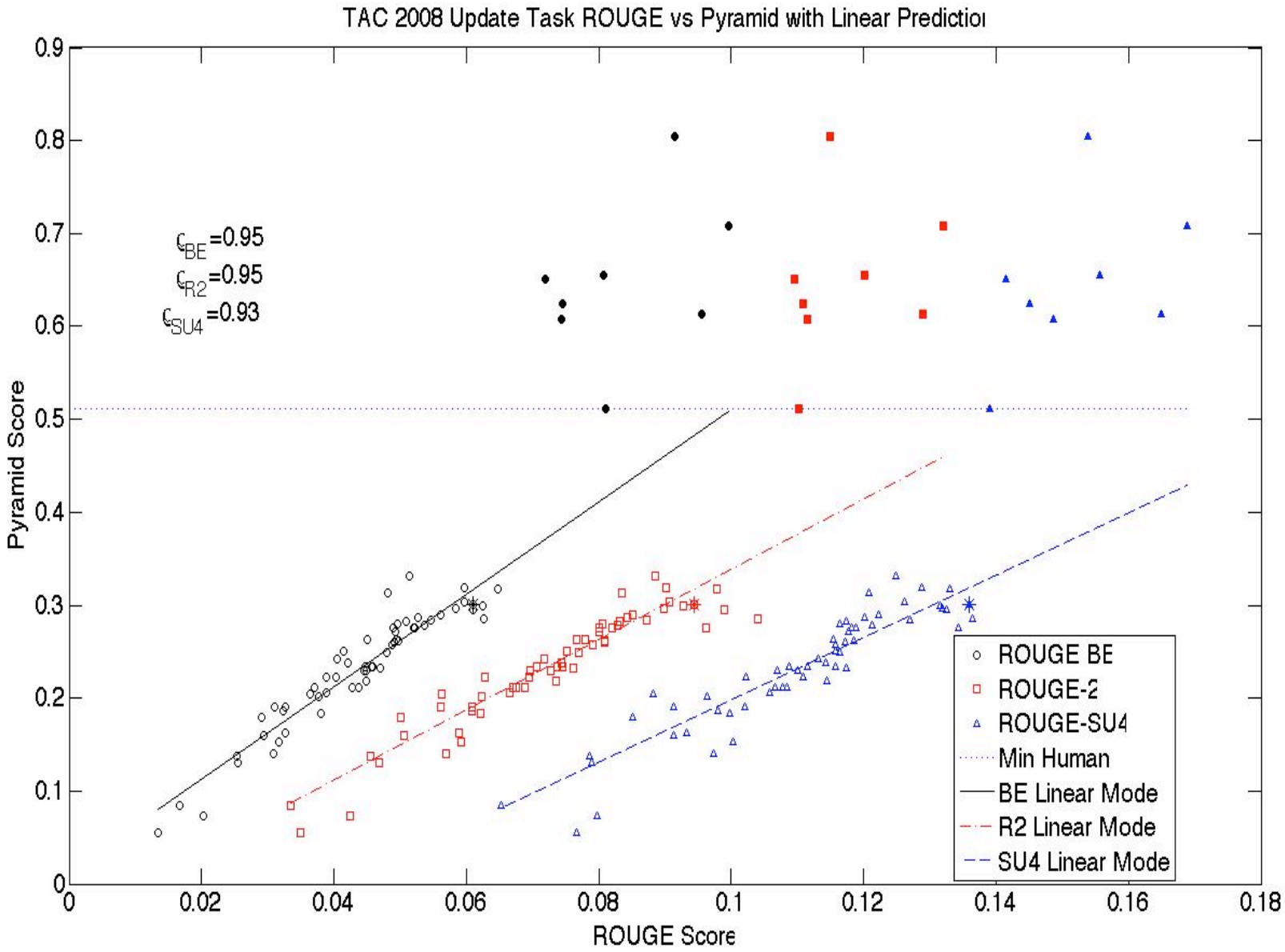
# Our *Favorite* Metric:ROUGE 1



# ROUGE and Responsiveness



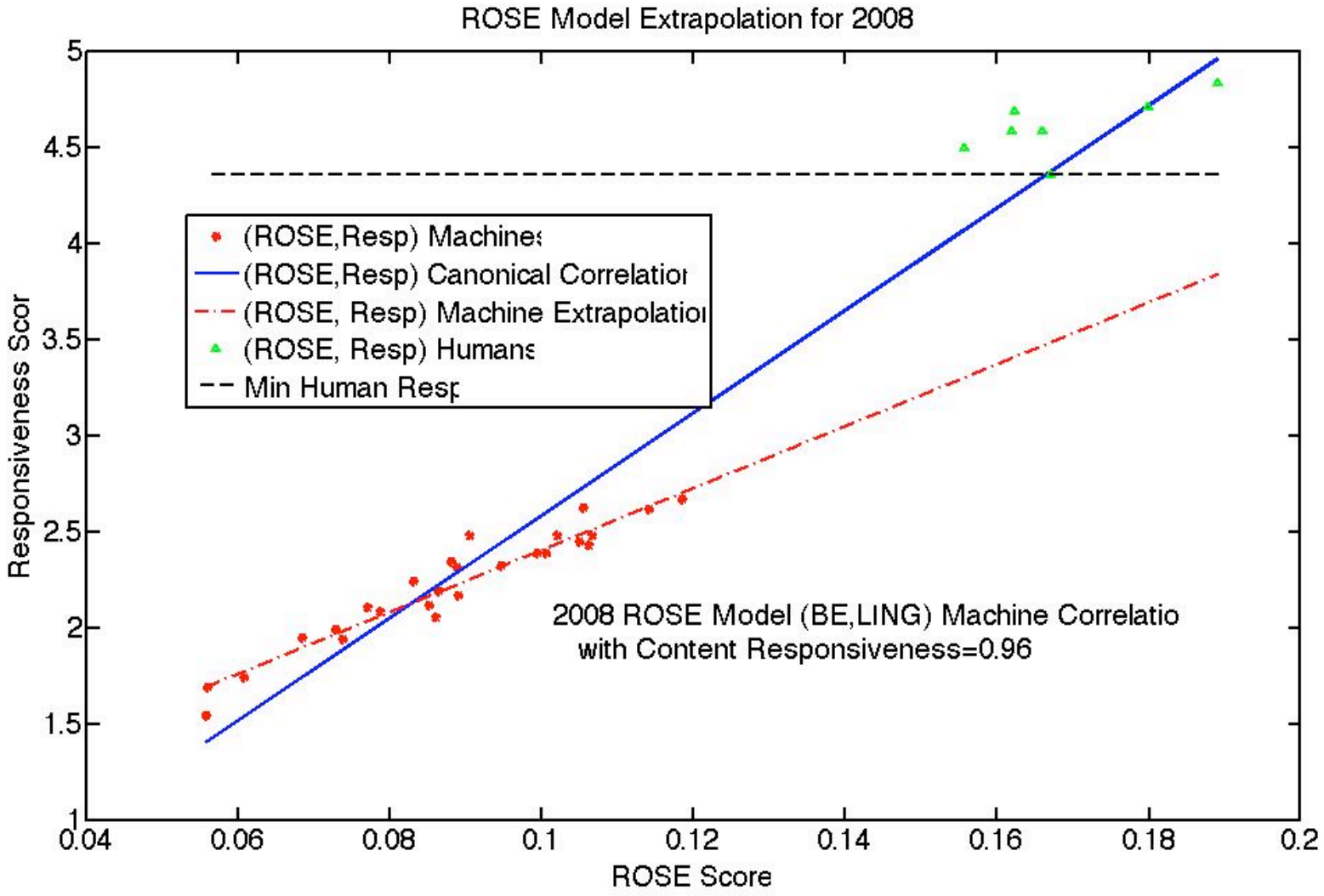
# Correlating ROUGE with Pyramid



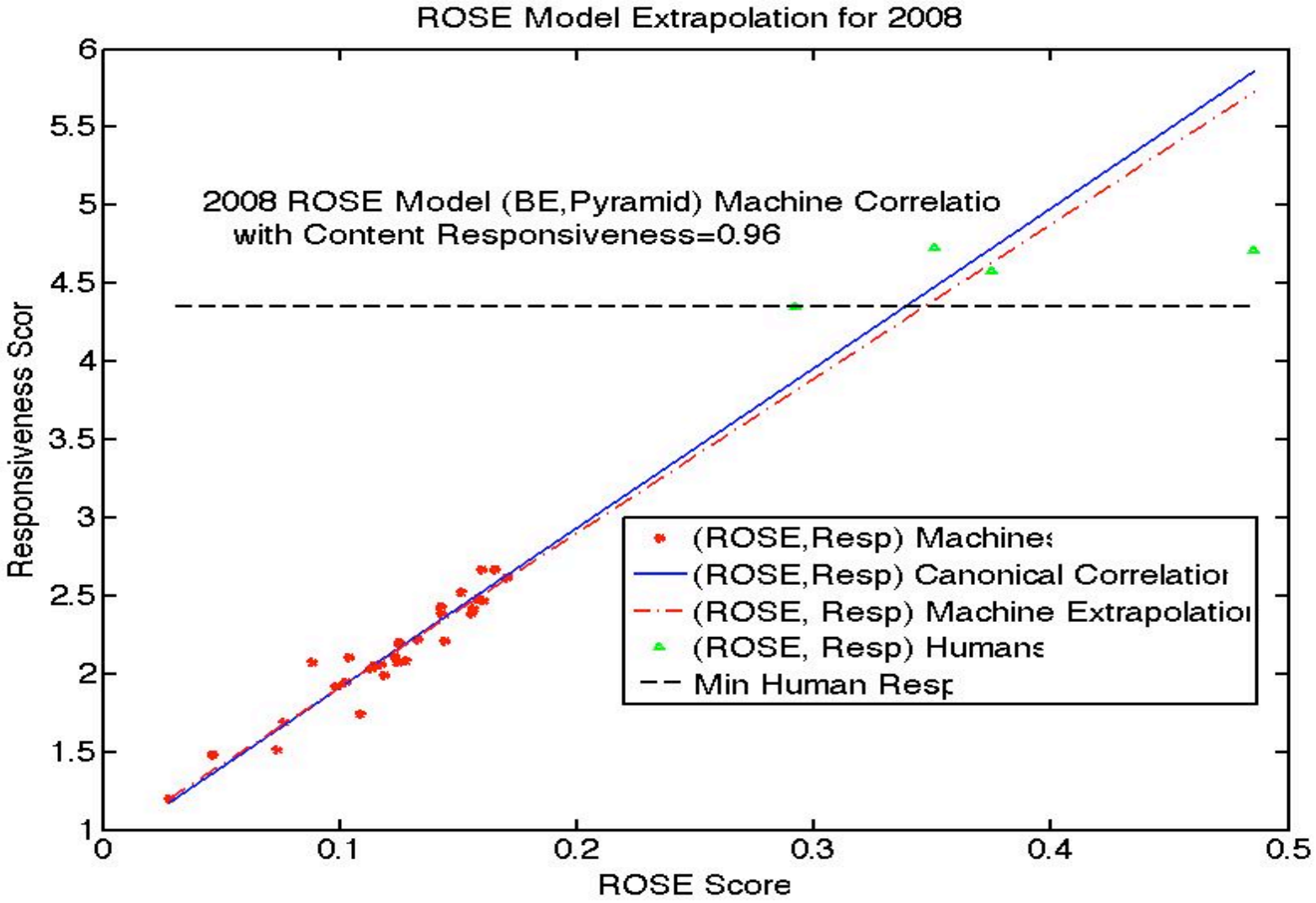
# Choose Best Linear Combination of Metrics

- Canonical Correlation: Hotelling 1935
  - Finds optimal linear combination to maximize correlation: a LS problem; more generally an eigenvalue problem.
- ROUGE Optimal Summarization Evaluation. ROSE, Conroy,Dang 2008.
- Linear combination of average system scores not document set scores.

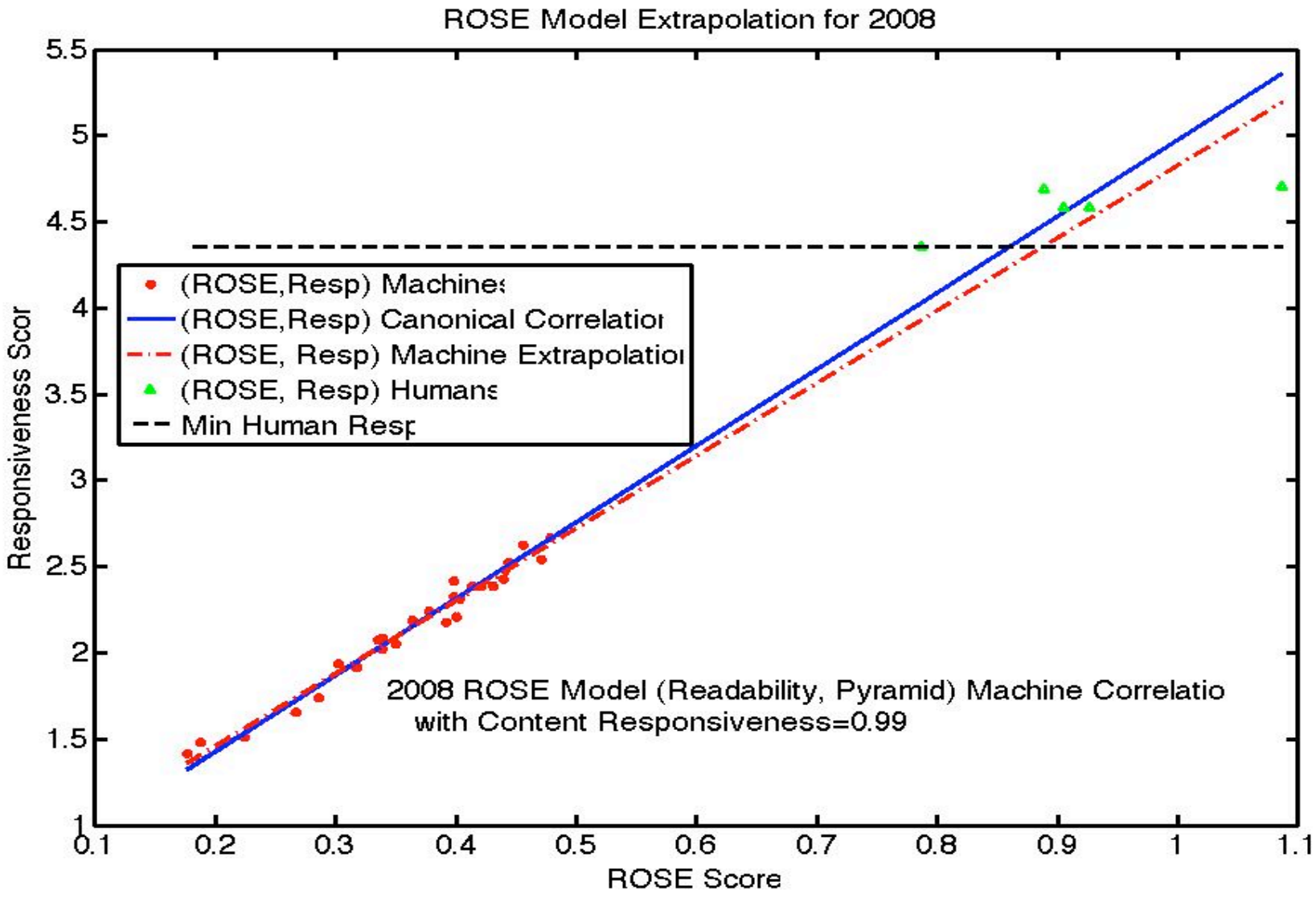
# (BE,Readability) Model



# (BE, Pyramid) Model



# (Readability, Pyramid) Model



# Conclusions

- CLASSY did well at ROUGE eval. for update task and on human evals.
- Gap between humans and machines still exists.
- Gaps automatic and human metrics still exists.
- Pyramid correlates quite well with overall responsiveness.

# Meta Evaluation

- Evaluate the Evaluation Methods.
  - Automatic methods to estimate:
    - Linguistic quality. (Regina Barzilay, Mirella Lapata 2005)
    - Pyramid scoring.  (Columbia, Univ. Penn.)
  - New ROUGE BE, n-gram graph evaluation.
  - Correlate overall responsiveness with an extrinsic evaluation: What task is the summary serving?

# Easy and Hard to Please