Bridging the ROUGE/Human Evaluation Gap in Multi-Document Summarization

John M. Conroy
Judith D. Schlesinger
IDA Center for Computing Sciences,USA
Dianne P. O'Leary
University of Maryland, College Park, USA

Outline

- CLASSY 07
  - Main: System 24.
  - Update: System 44.
- Gaps in performance and metrics.
- Comparison MSE 2006. (panel tomorrow)
- Better metrics? (panel tomorrow)

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

- Linguistic preprocessing.
  - Shallow parsing
  - Find sentences and shorten them.
- Sentence Scoring.
  - Approximate Oracle.
- Redundancy Removal.
  - Select a subset of sentences.
  - LSI and L1-norm QR.
- Ordering
  - TSP

Processing: Structure and Linguistic

- Use sgml tags to remove datelines, bylines, and harvest headlines.
- Use heuristic patterns to find phrases/clauses/words to eliminate
  - Finding sentence boundaries.
  - Shallow processing.
- Removed lead pronoun sentences and question sentence for 2007.

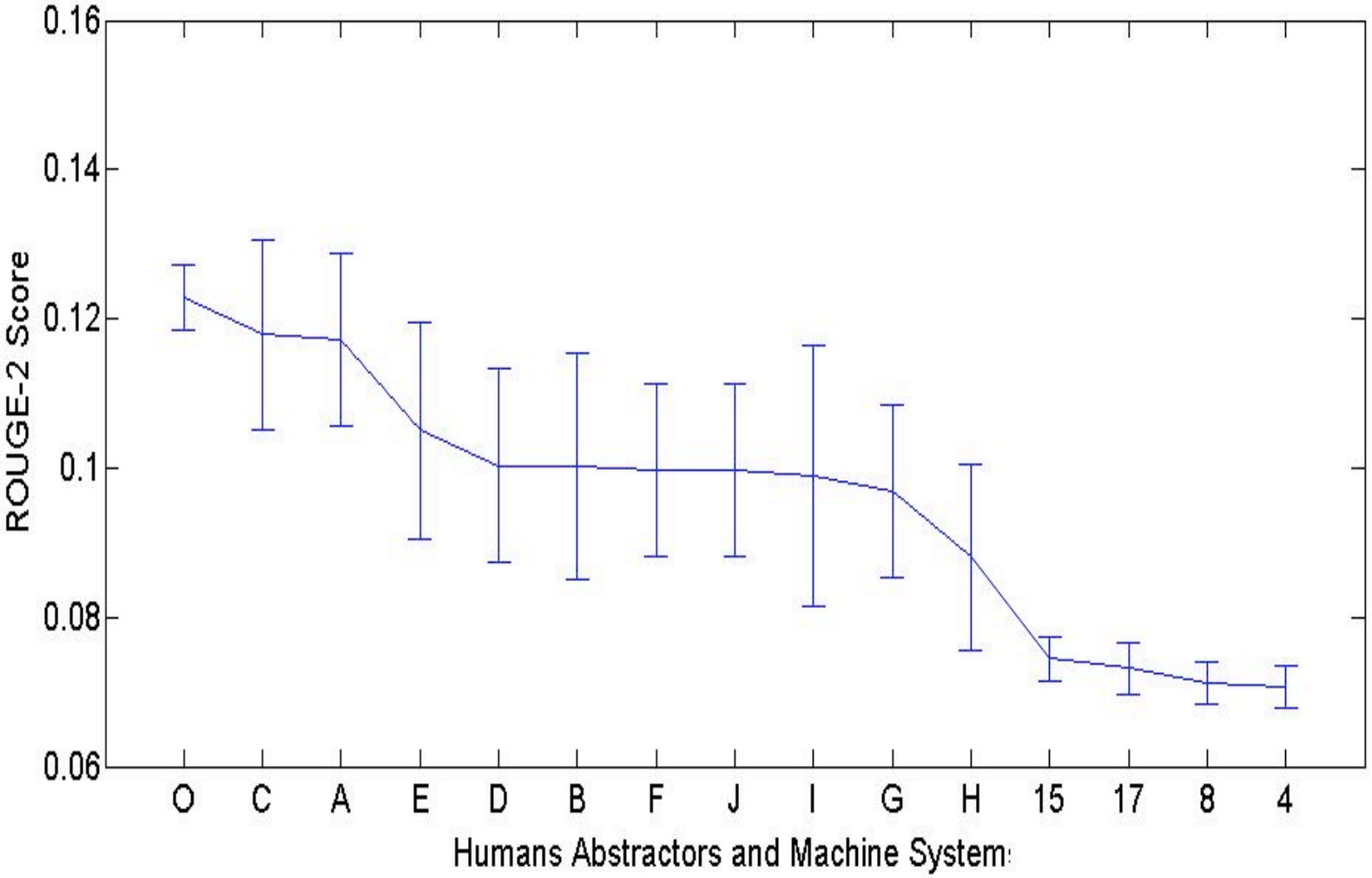
Linguistic Processing

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

An Oracle and Average Jo

- 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.
  - "Average Jo" Oracle Score: fraction of expected abstract terms (vector space model).

The Oracle Pleases Everyone!



Signature Terms

- Term: stemmed (lemmatized), space-delimited string of characters from {a,b,c,…,z}, after text is lower cased and all other characters and stop words are NOT removed.
- Need to restrict our attention to indicative terms (*signature terms*).
  - Terms that occur more often then expected.

Signature Terms

Terms that occur more often than expected in *Aquaint* collection.

- Based on a 2×2 contingency table of relevance counts.
- Log-likelihood; equivalent to mutual information.
- Dunning 1993, Hovy Lin 2000.

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

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

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

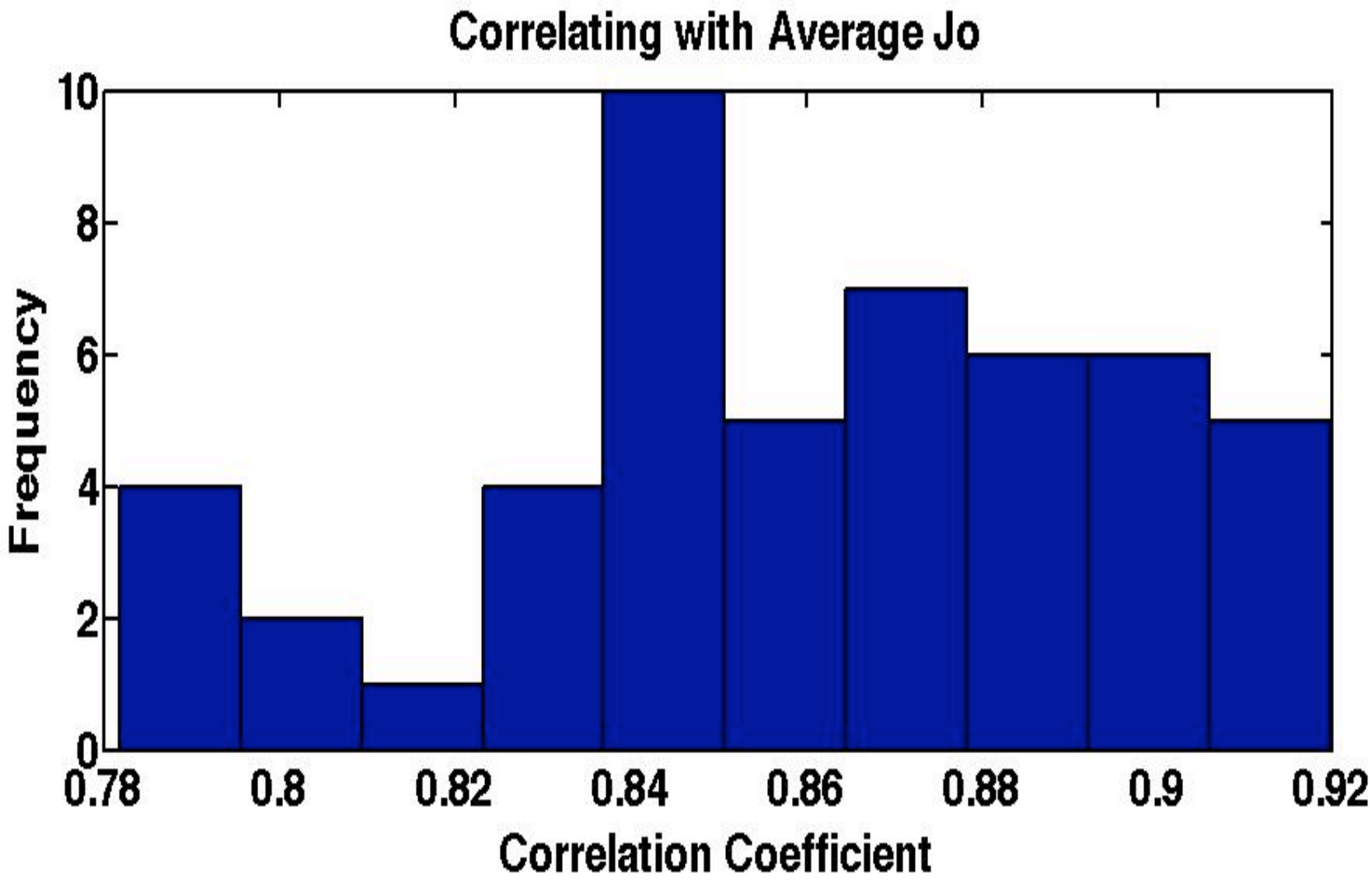
$$s(t) = \begin{cases} 1 \text{ if } t \text{ is a signature term} \\ 0 \text{ if } t \text{ is not a signature term} \end{cases}$$

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

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

- The score of a sentence is the sum of Pr(t) taken over its terms divided by its length.

Correlation with Oracle



Smoothing and Redundancy Removal

Use approximate oracle to select candidate sentences (~750*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 0.5*n*.

– L1 pivoted QR to select sentences.

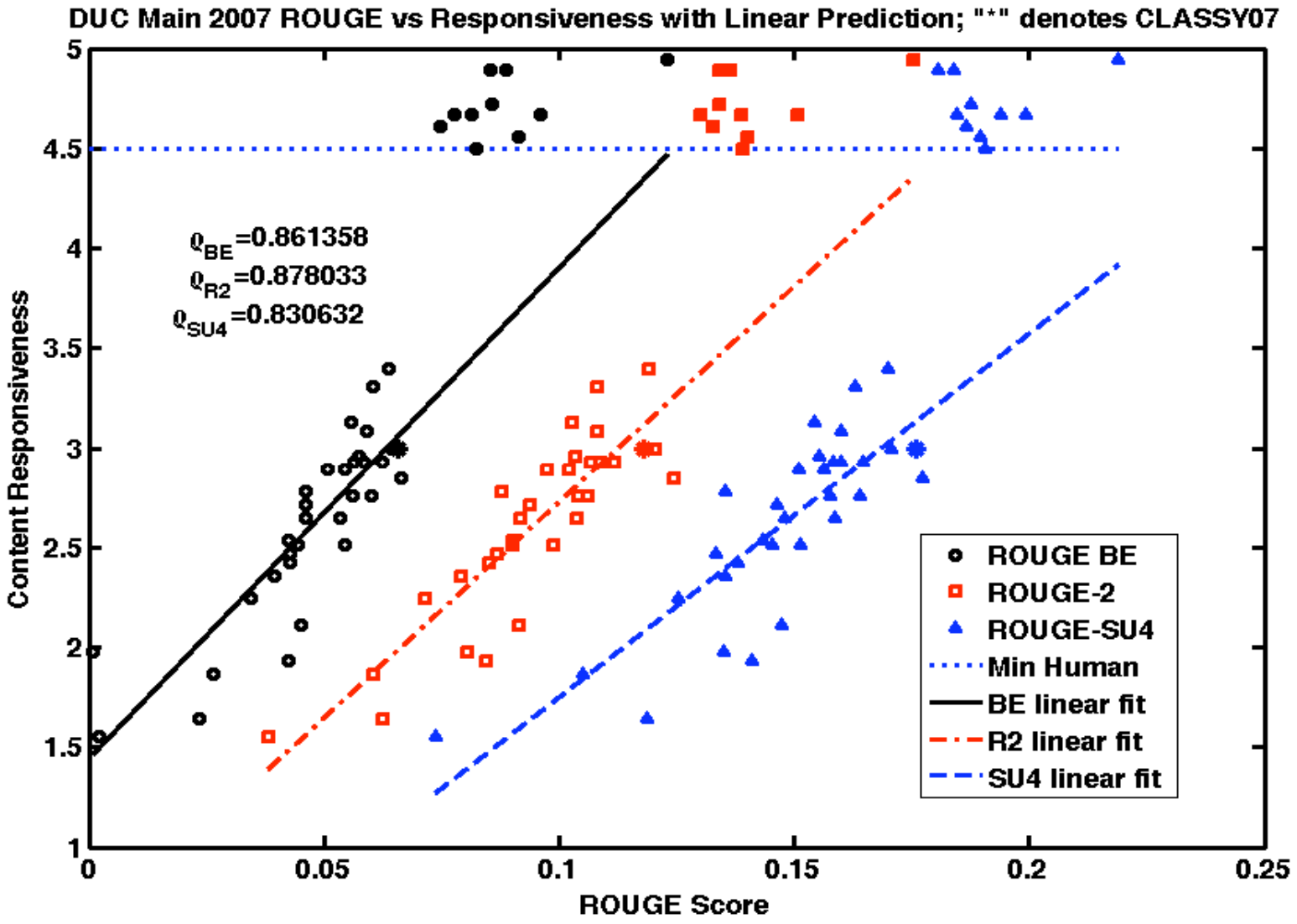
	s_1	$\cdots$	s_n
t_1	a_{11}	$\cdots$	a_{1n}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_m	a_{m1}	$\cdots$	a_{mn}

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.

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

DUC 2007: Main Task



Why the Gap?

- Should Evaluators=Human Summarizer?
- Advantage:
  - Person writing summary judges all summaries?
- Disadvantage:
  - Personal interest (bias?) affects assessment.
- Mean Human score DUC 07 was 4.9.
  - Removing self assessment score was 4.7, T-test indicates humans like their own summary more than other human summaries.
- Do we aim to target every human's ideal or find a middle ground (ROUGE) to please the masses? Come to the panel discussion…

Linguistics vs. Responsiveness

- Evaluators liked summaries ending with a period. [Lucy] ($2.8 \neq 2.5$ with 96% conf).
- But, no significant difference in ROUGE-2.
- Responsiveness in DUC 07 was suppose to be content only and not overall.
- However,…

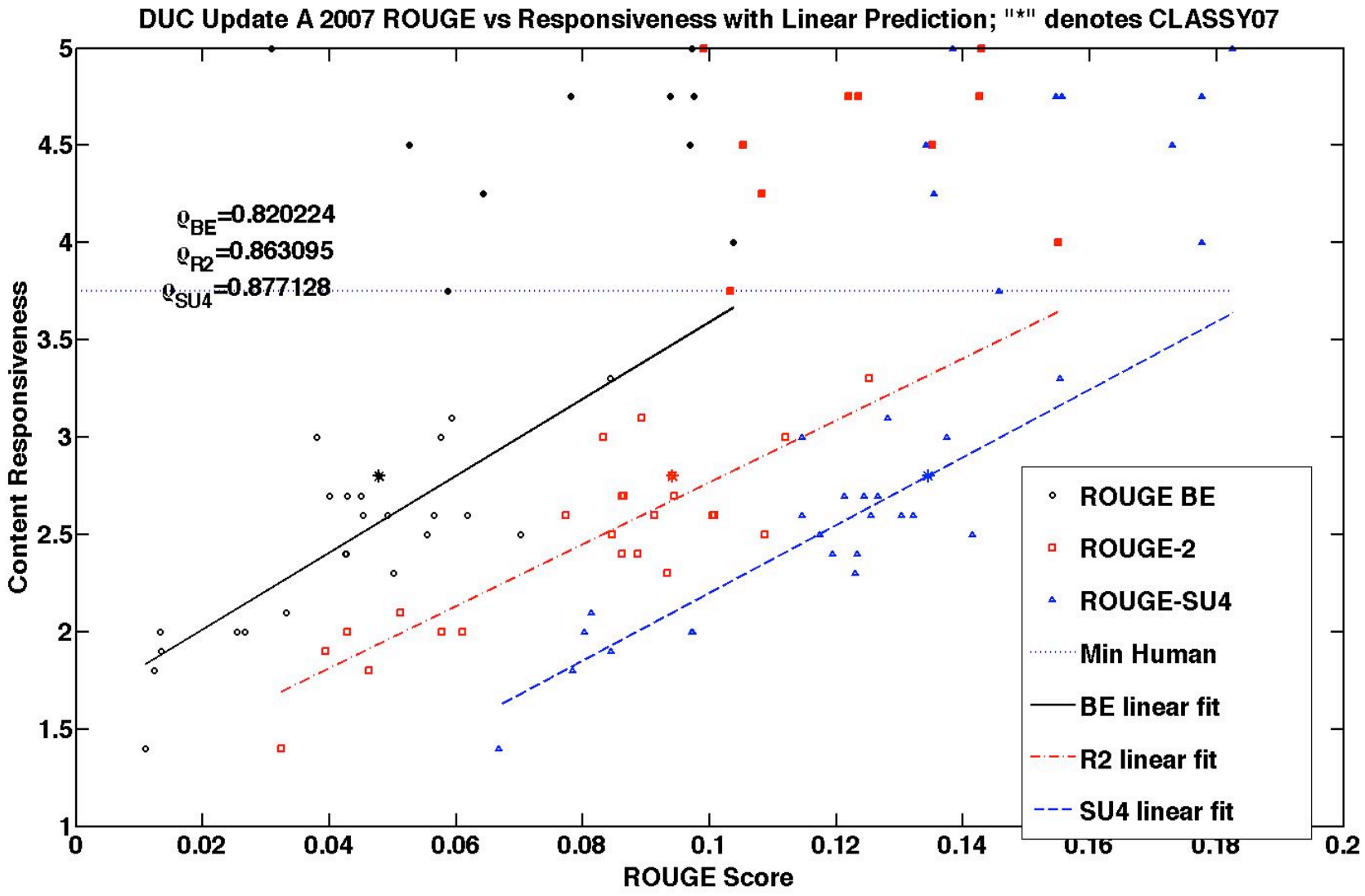
Correlating Linguistics Responsiveness

Question	Content Resp. 06	Overall Resp. 06	Content Resp. 07
Grammar	0.32	0.50	0.60
Non-Red.	-0.37	-0.24	-0.43
Ref. Clarity	0.24	0.53	0.59
Focus	0.39	0.62	0.71
Structure Coherence	0.13	0.46	0.49

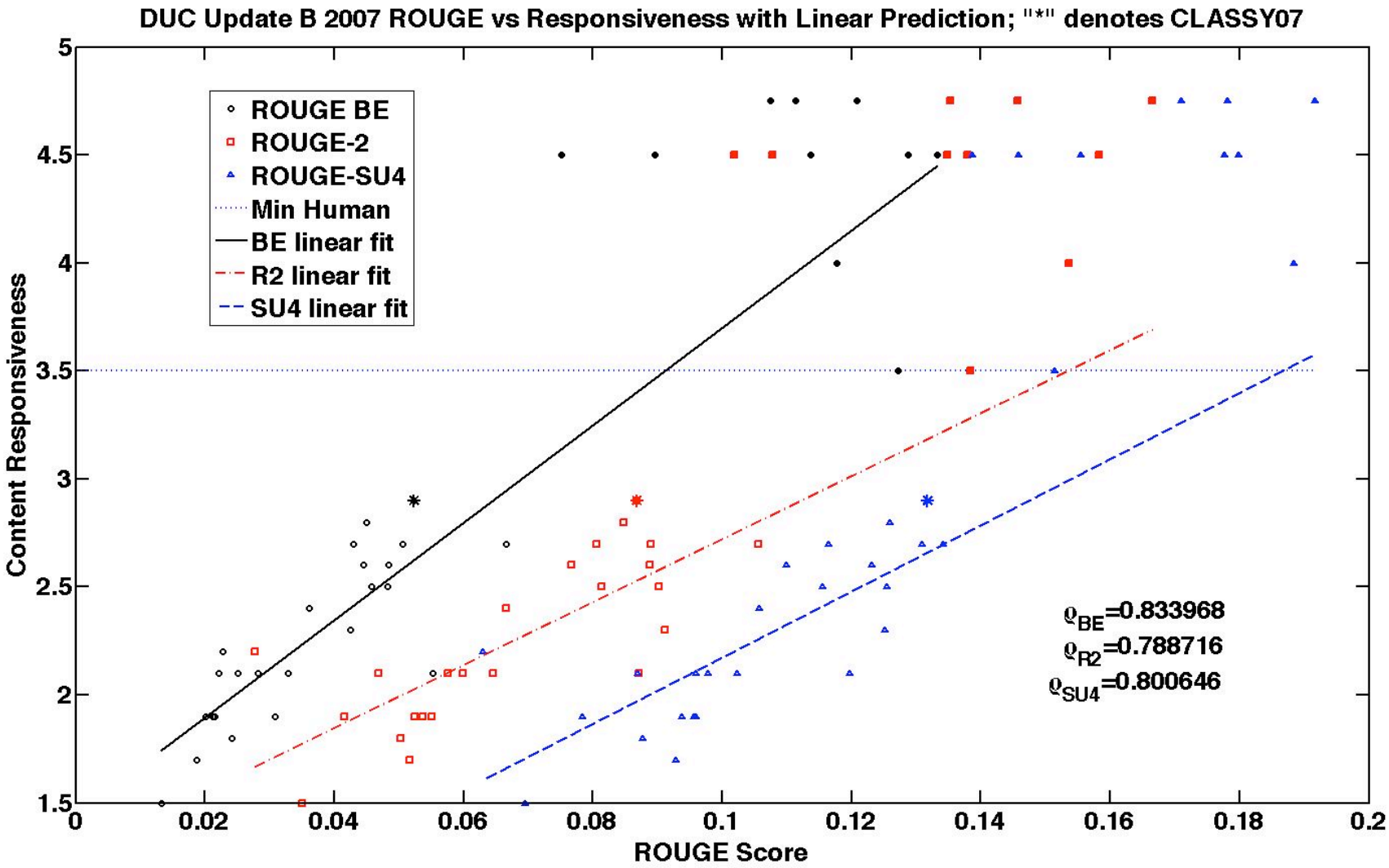
Adaptations for Update

- Sub-task A: run CLASSY 07 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 8 new documents.
- Sub-task C: analogous to sub-task B submission.

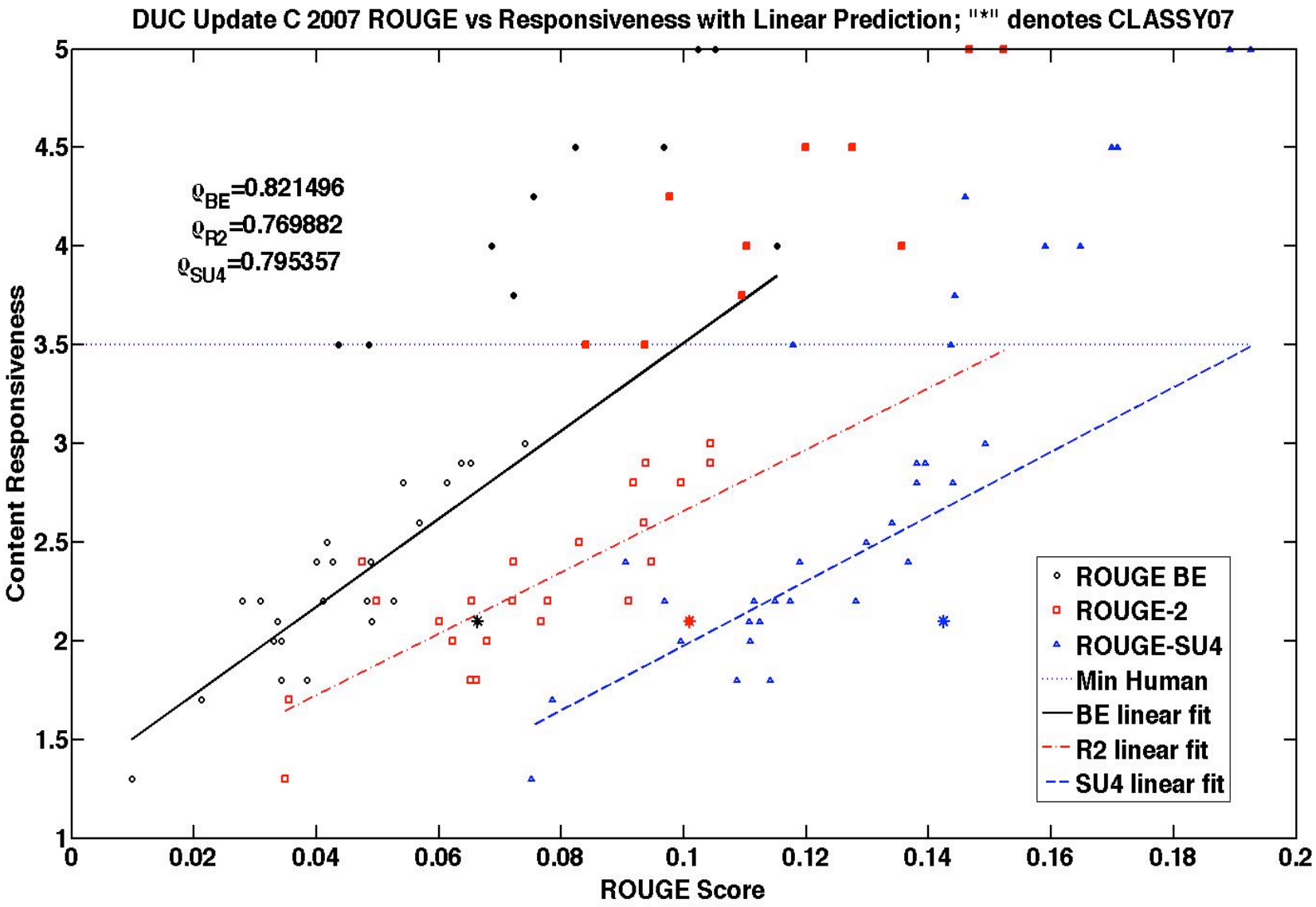
Update: Sub-task A



Update: Sub-task B



Update: Sub-task C



Conclusions

- CLASSY 07's did extremely well at ROUGE evaluation for main task and well on human eval.
- Gap between humans and machines still exists.
- Gap between ROUGE and responsiveness still exists.
- Both human and automatic evaluation should be rethought. (Stay tuned for panel discussion, tomorrow.)
- Looking forward to more update evaluation.