AESOP: Summarization and Metrics With Neither Sweet *Lemons* nor Sour *Grapes*

John M. Conroy

Judith D. Schlesinger

IDA Center for Computing Sciences, USA

Dianne P. O'Leary, University of Maryland

Outline

- Content based metrics
  - ROSE (ROUGE Optimal Summarization Evaluation).
  - Nouveau ROUGE: measuring what’s new.
- AESOP results.
- Uber-baseline: Towards automatic measures of coherence.

Best Linear Combination

- 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.

Robust Regression

- We aim to predict human metrics:
  - – Overall responsiveness or
  - – Pyramid evaluation.

$$x = \arg\min \| Ax - b \|$$

A_{2008} system-average-scaled-feature matrix,

b_{2008} is the human metric to predict,

$\|.\|$ a norm that accounts for outliers.

$\hat{b}_{2009} = A_{2009}x$, our estimate for the 2009 metric.

Nouveau ROUGE: Newness Metrics

- For update summaries the summaries should differ from what is already known.
- ROUGE scores that compare peers in subset B with models in subset A.

$$R_i^{(AB)} \quad i = 1,2,3,4,5,\mathrm{SU}4,L$$

Classifier

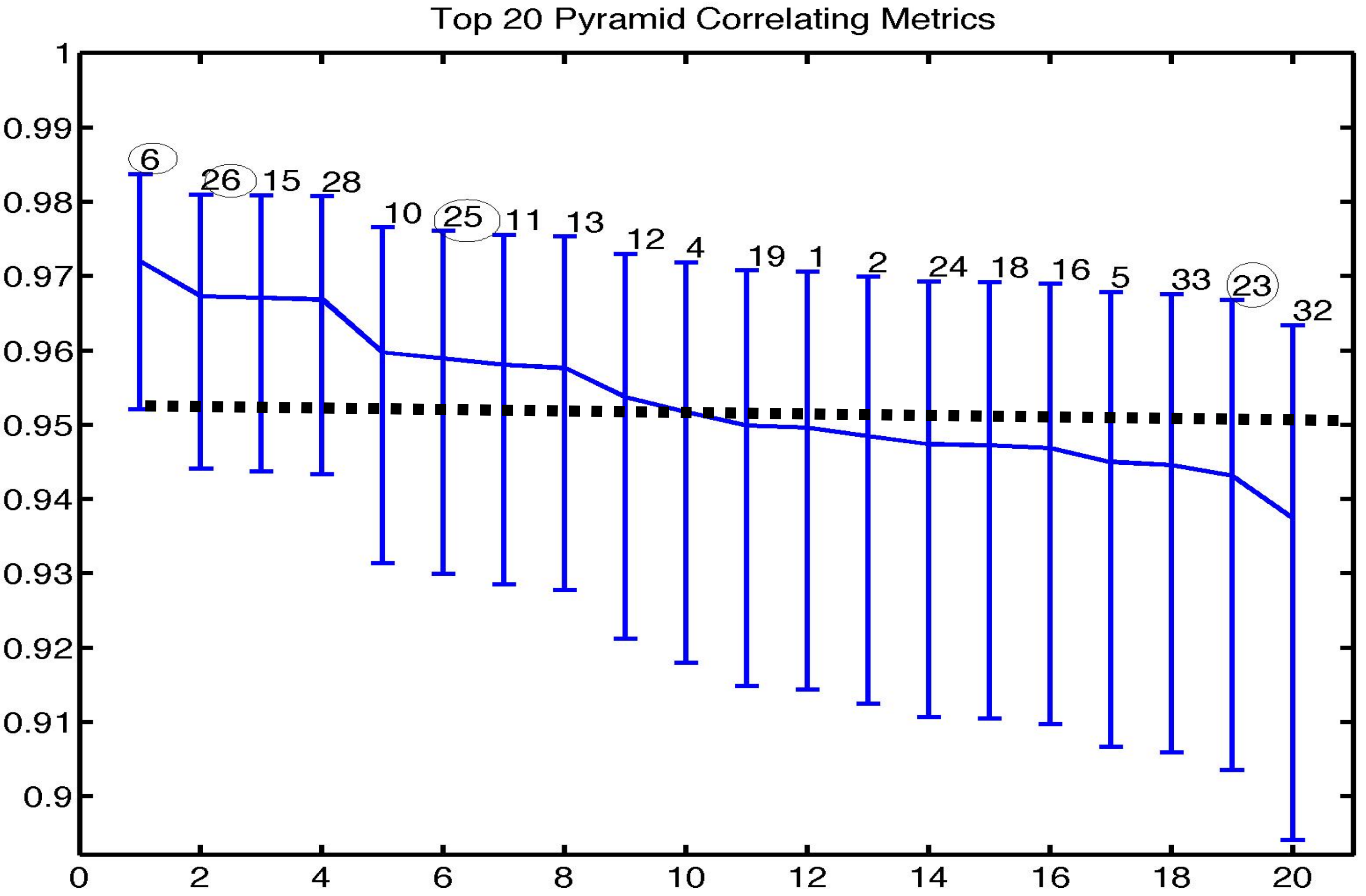
- Predict 2009 *document set* responsiveness scores using a linear classifier with ROUGE [and Nouveau ROUGE] features.
- Responsiveness scores for 2008 are {1,2,3,4,5}.
- Classifier gives posterior probability for each class.
- Expected value computed as score:

$$s = \sum_{i=1}^{5} i p_i$$

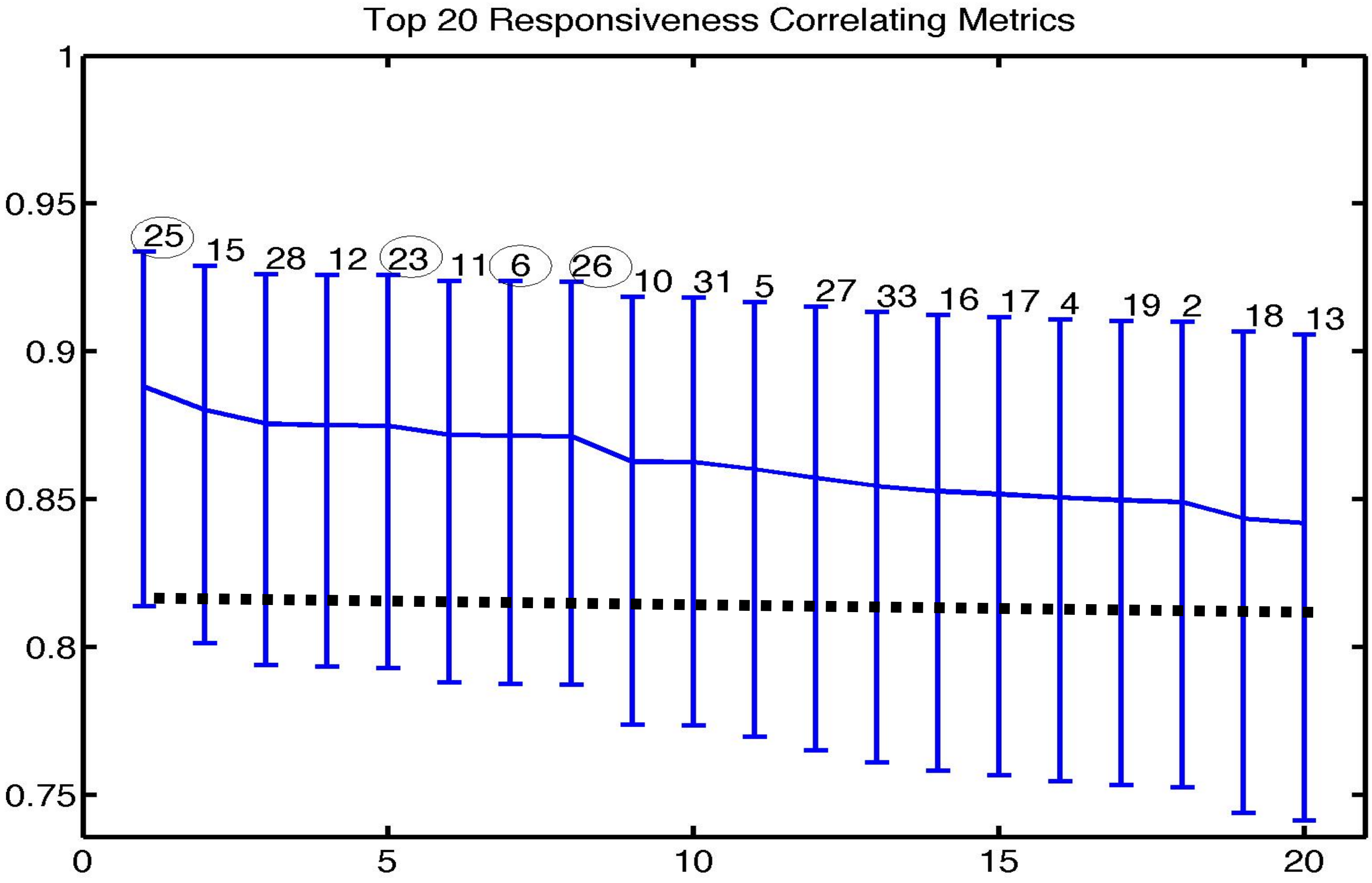
AESOP Submissions

ID	Type	Features	Target
25	Regress.	1,2,3,L,SU4	Resp.
6	Regress.	2	Resp.
23	Regress.	1,2,3,L,SU4	Pyramid
26	Classifier	2,3	Resp.

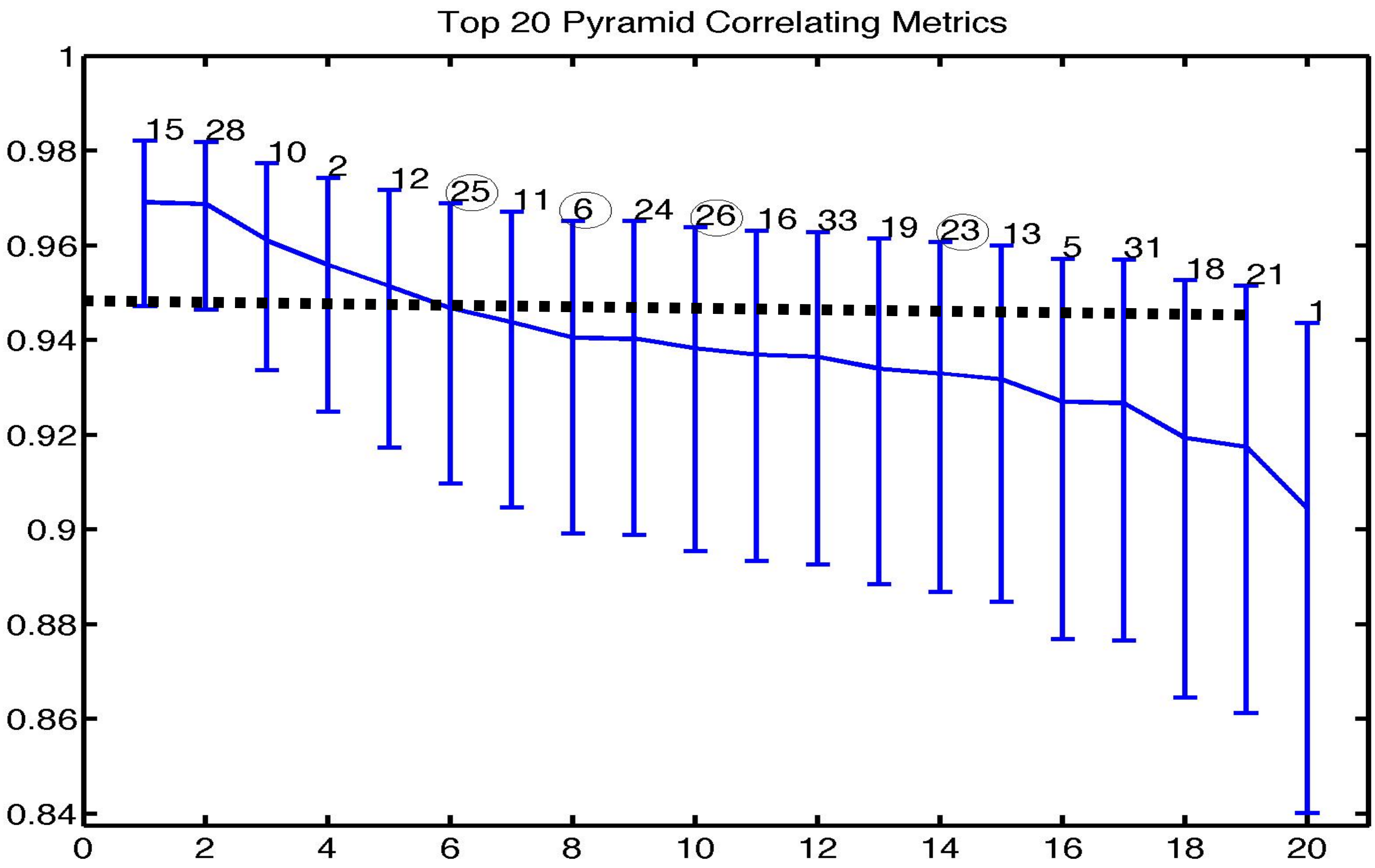
Pvramid Set A: Error Bars



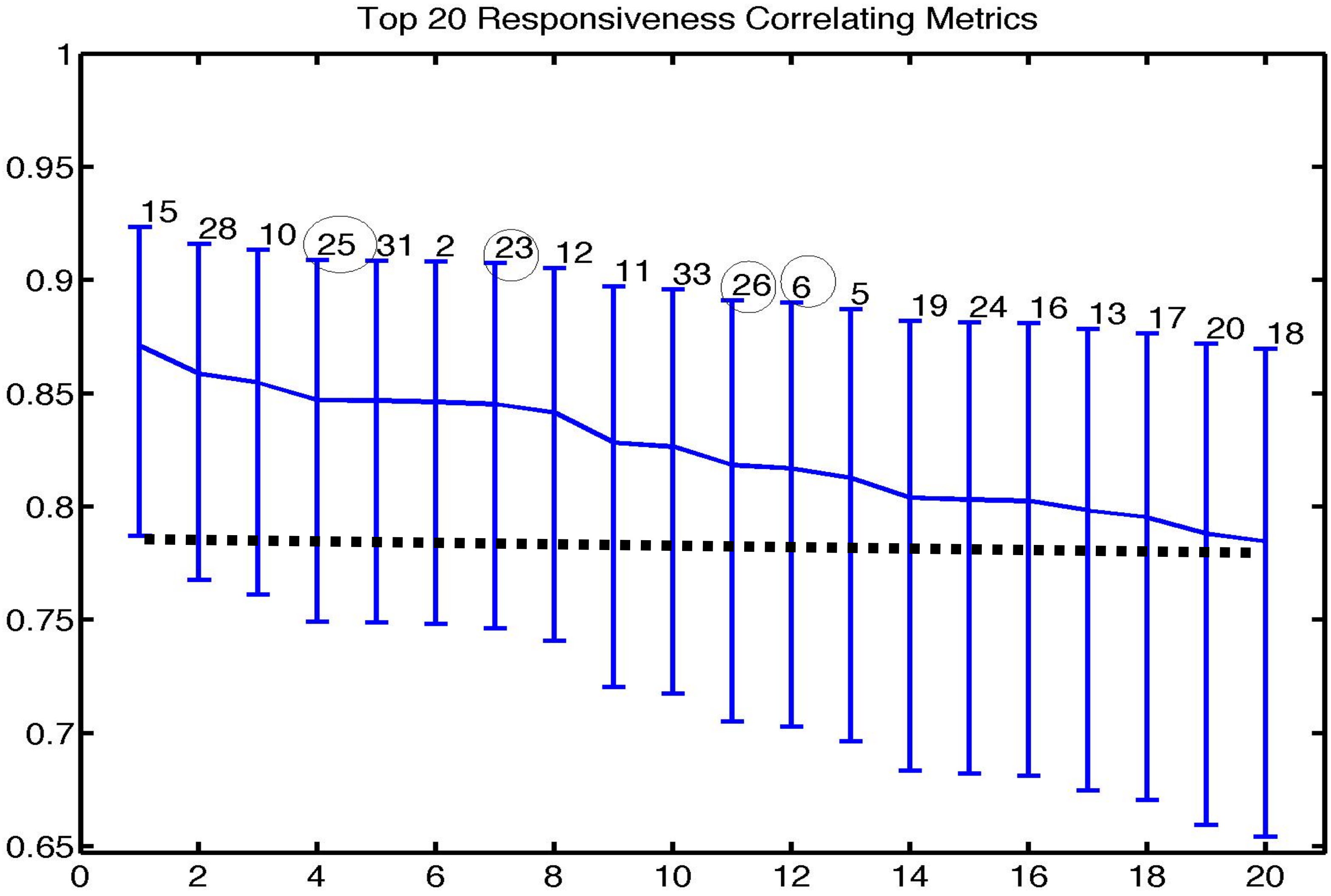
Responsiveness Set A



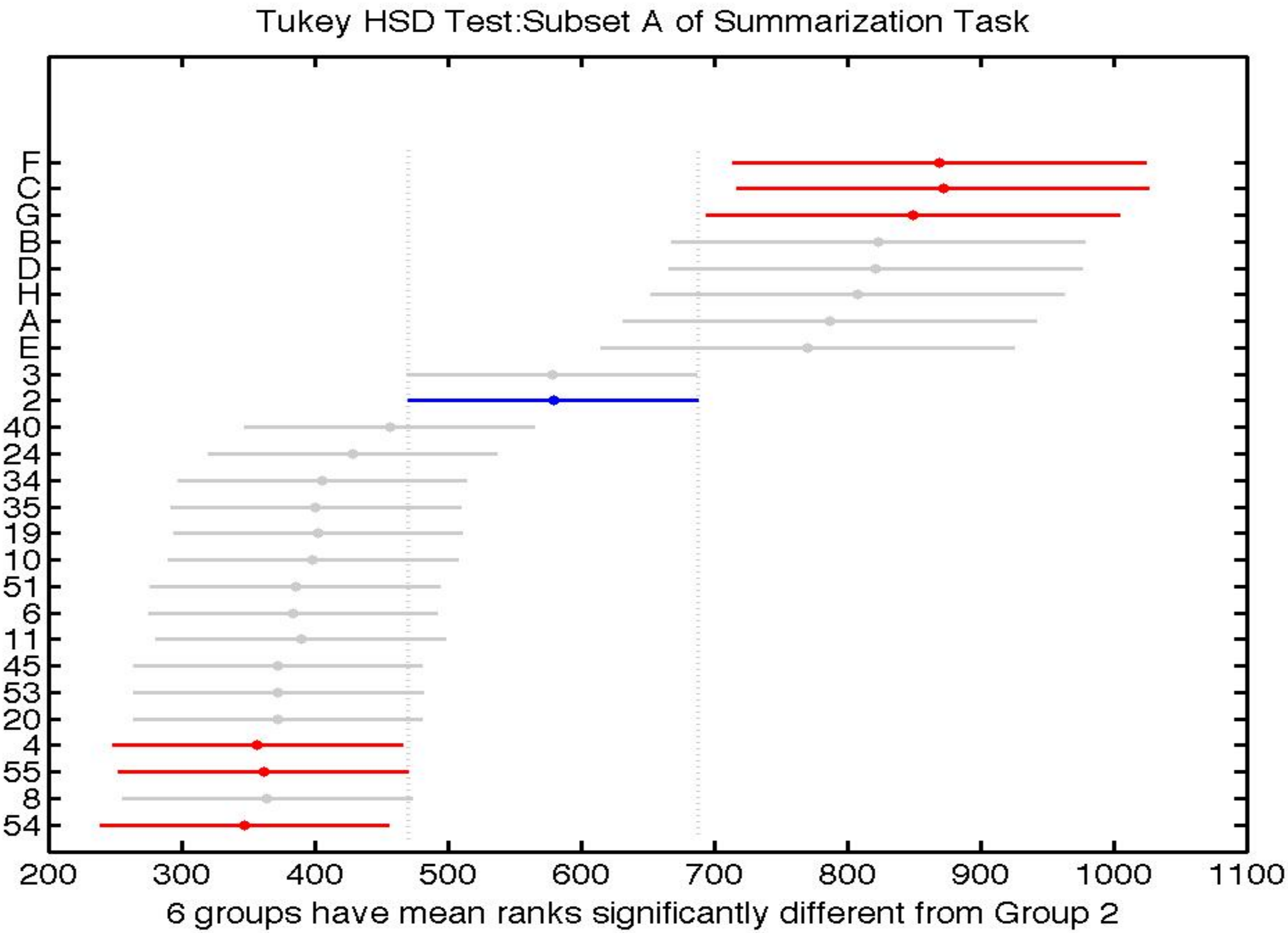
Pyramid Set B: Error Bars



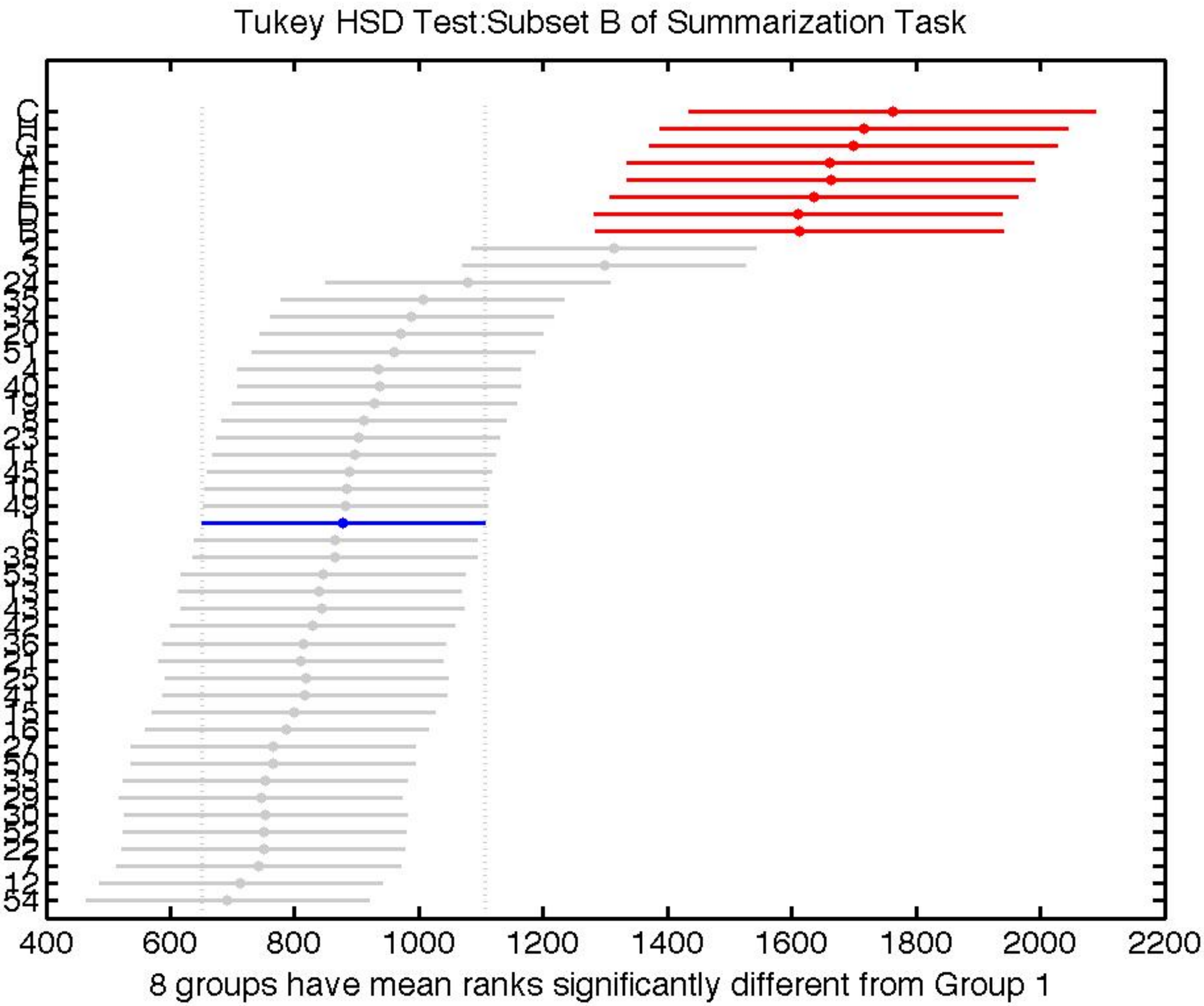
Responsiveness Set B



Responsiveness: Set A



Responsiveness: Set B



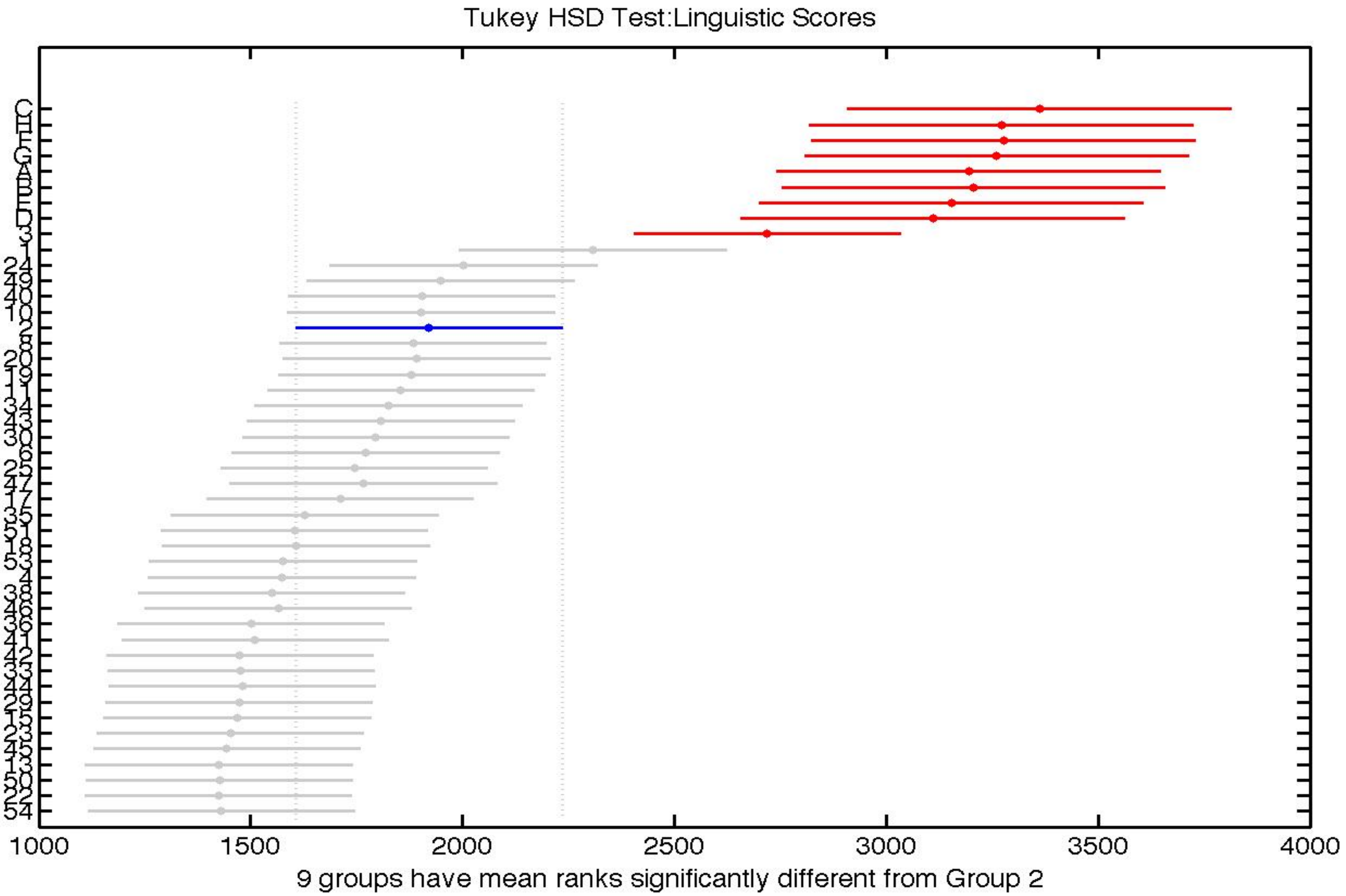
Uber-Baseline

- Idea: Test to what extent sentence order affects linguistic quality and responsiveness.
- Execution: Permute sentences from a human summary (not the assessor for the topic set.)

Metrics on the Uber-Baseline

Metric	Uber	Human	p-value
pyr	0.656	0.662	9.40e-01
ling	5.682	8.773	5.92e-14
overall	6.273	8.591	6.04e-13

Uber vs The Top



Conclusions

- While ROSE/Nouveau ROUGE and others had higher correlation than baseline metrics, none exceeded ROUGE-2 for predicting responsiveness.
- Linguistic quality of uber-baselines comparable to top performing systems; however, *significantly* less than human counterpart!
- Underscores need for coherence metrics.