Crafting Strong Predictors of a Summary's Quality: "Essentially, all models are wrong, but some are useful"*

John M. Conroy[1]

Peter A. Rankel[2]

Judith D. Schlesinger[1]

Dianne P. O'Leary[2]

1. IDA Center for Computing Sciences, USA

2. University of Maryland

*George E. P. Box, Univ. Wisconsin, Prof. Emeritis

Overview

- Our Models
  - Prediction via regression or eigenvectors
  - Features:
    - Content
    - Nouveau
    - Linguistic
  - Feature Selection

Modeling Background

- Canonical Correlation: Harold 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]
- Nouveau-ROUGE, [Conroy, Schlesinger, O'Leary, Computational Linguistics 2011]
- Linear combination of *average system scores* *not* document set scores.

Robust Regression and Non-Negative Least Squares

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

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

$A = A_{2009}$ system-average-feature matrix,

$b = b_{2009}$ is the human metric to predict,

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

$\hat{b}_{2010} = A_{2010}x$, our estimate for the 2010 metric.

Optionally, we can add contraint of non-negativity of x.

Canonical Correlation

- Find a linear combination of features and a linear combination of human judgments (pyramid, resp., ling.) with highest correlation.

$$(x, y) = \arg\max_{x \in R^n, y \in R^3} \rho(Ax, By)$$

- Where $\rho()$ is Pearson correlation.
  - Training is solving a generalized eigenvalue problem.
  - Score using *x* only.

Content Features and Newness Features (Nouveau-ROUGE)

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

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

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

Linguistic Features: One Matrix and 7 Features

1. Log sum term overlap between consecutive sentences *(L_{o1})*
2. Summary normalized term overlap *(L_{o2})*
3. Redundancy Score 1:dist. to rank 1 *(L_{r1})*
4. Redundancy Score 2:dist. to rank 2 *(L_{r2})*
5. -log(number of sentences) *(L_{sl})*
6. Term Entropy *(L_{te})*
7. Sentence Entropy *(L_{se})*

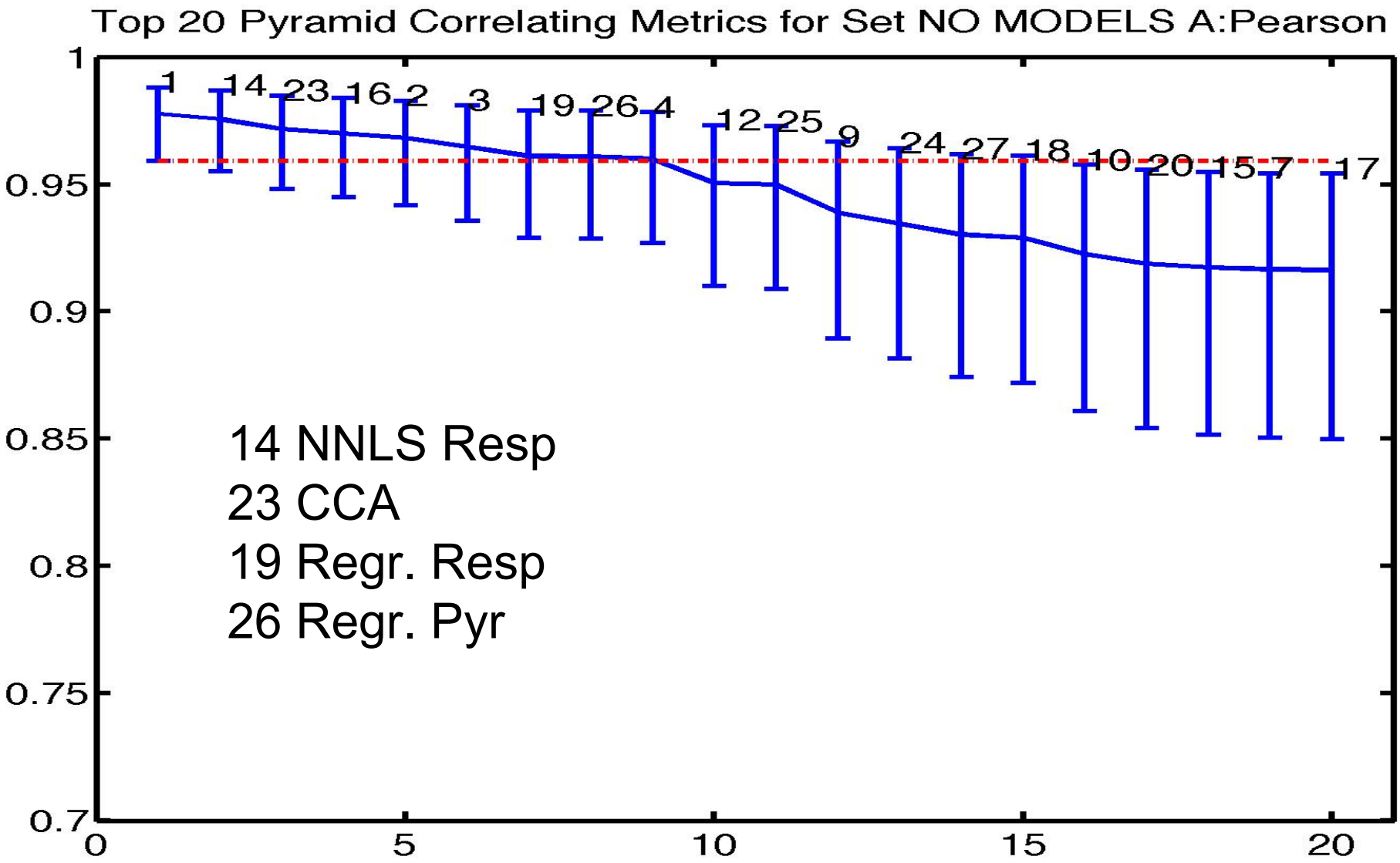
Training Models

- Feature Selection:
  - Train 2^{14}-1 models all proper subsets of 14 features computed from TAC 2008.
  - Evaluate (best correlation) on TAC 2009.
  - (Repeat for update set)
- Train best models on TAC 2009 and score for TAC 2010.

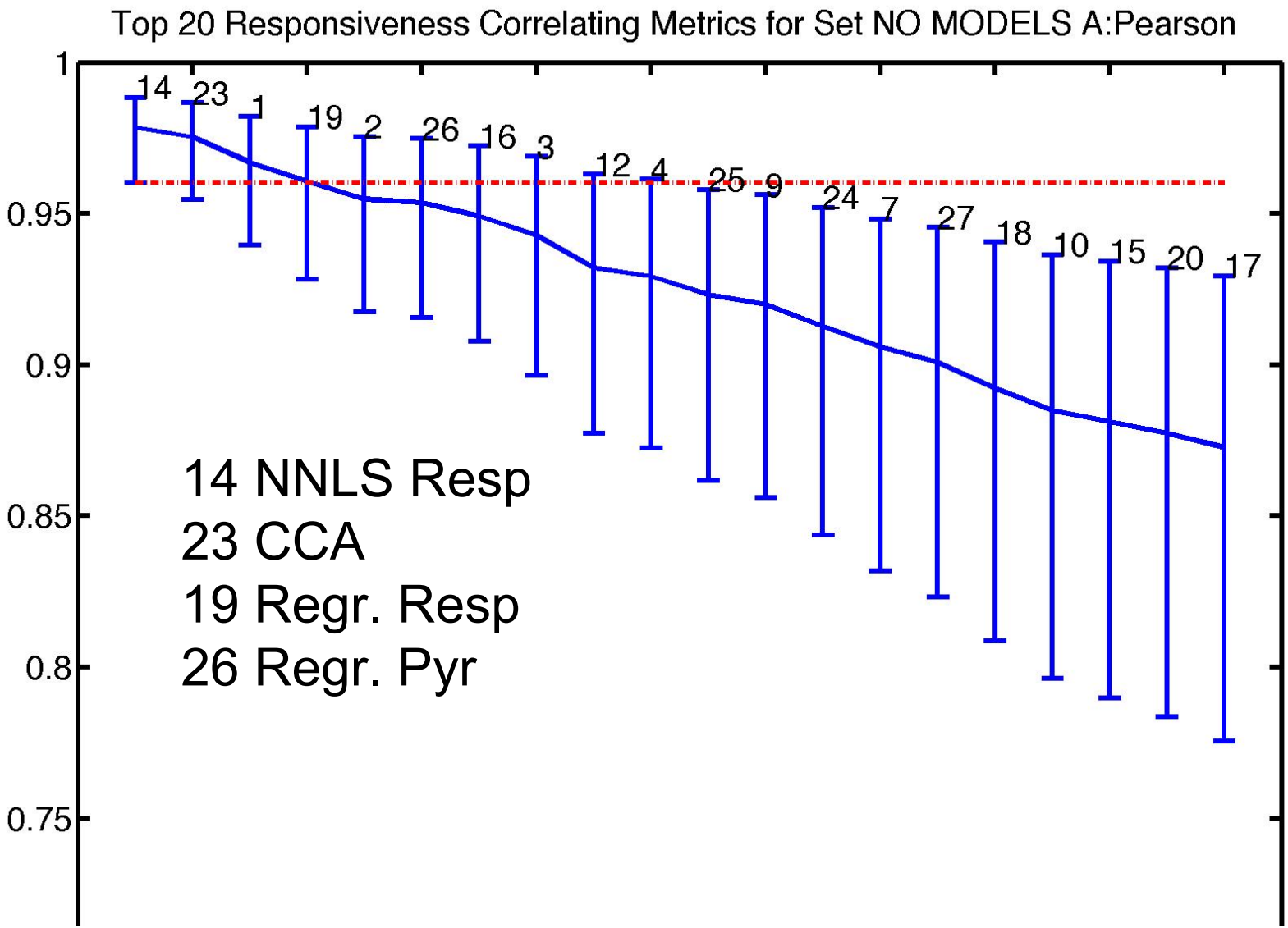
AESOP Submissions:No Models

ID	Type	Features	Target
14	NNLS	R_2, L_{o1}, L_{sl} R_2, R_5, R_L, L_{o1}	Resp.
23	CCA	$R_2, R_L, L_{o1}, L_{o2}, L_{r1}, L_{sl}, L_{se}$ $R_1, R_3, R_4, R_{SU4}, L_{o1}, L_{r1}, L_{r2}, L_{sl}, L_{se}$	Resp., Pyr., Ling.
19	Robust Reg.	$R_1, R_2, R_{SU4}, L_{r2}, L_{sl}, L_{se}$ $R_2, L_{o2}, L_{r1}, L_{r2}, L_{sl}, L_{se}$	Resp.
26	Robust Reg.	$R_1, R_2, R_{SU4}, L_{o1}$ $R_2, R_3, R_{SU4}, L_{o1}$	Pyramid

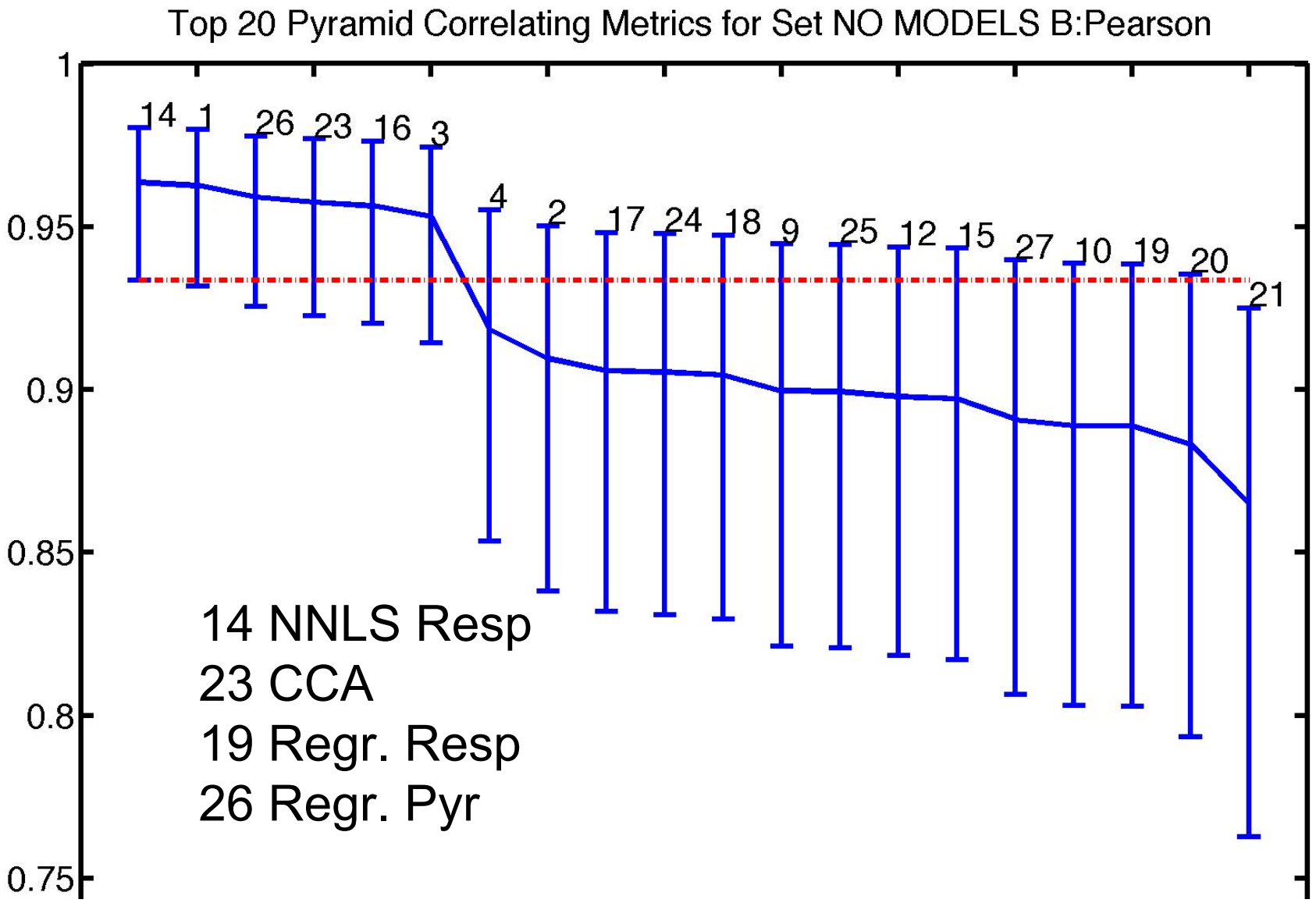
Pyramid Initial (A): Error Bars



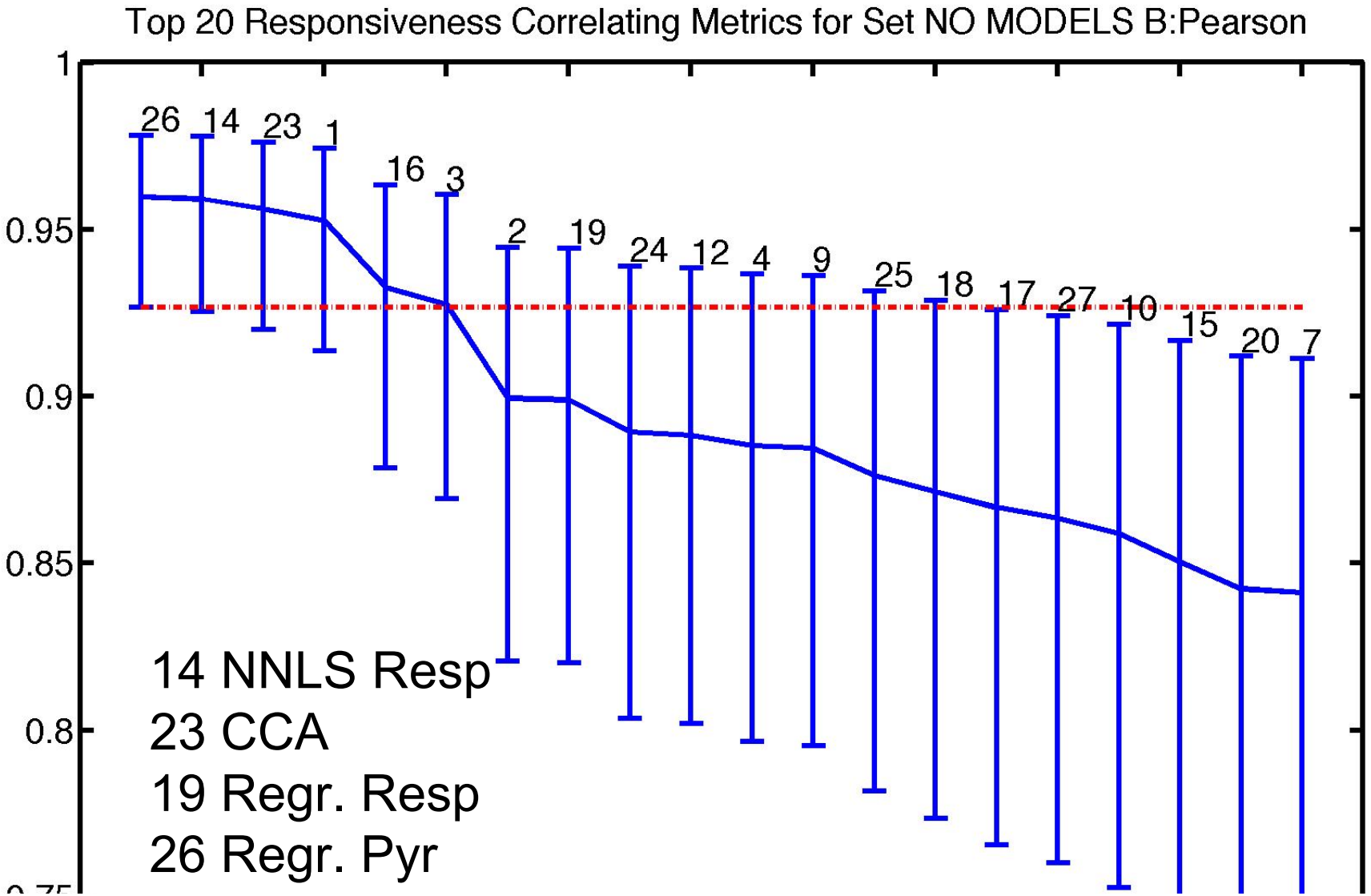
Responsiveness: Initial (A)



Pyramid: Update (Set B)



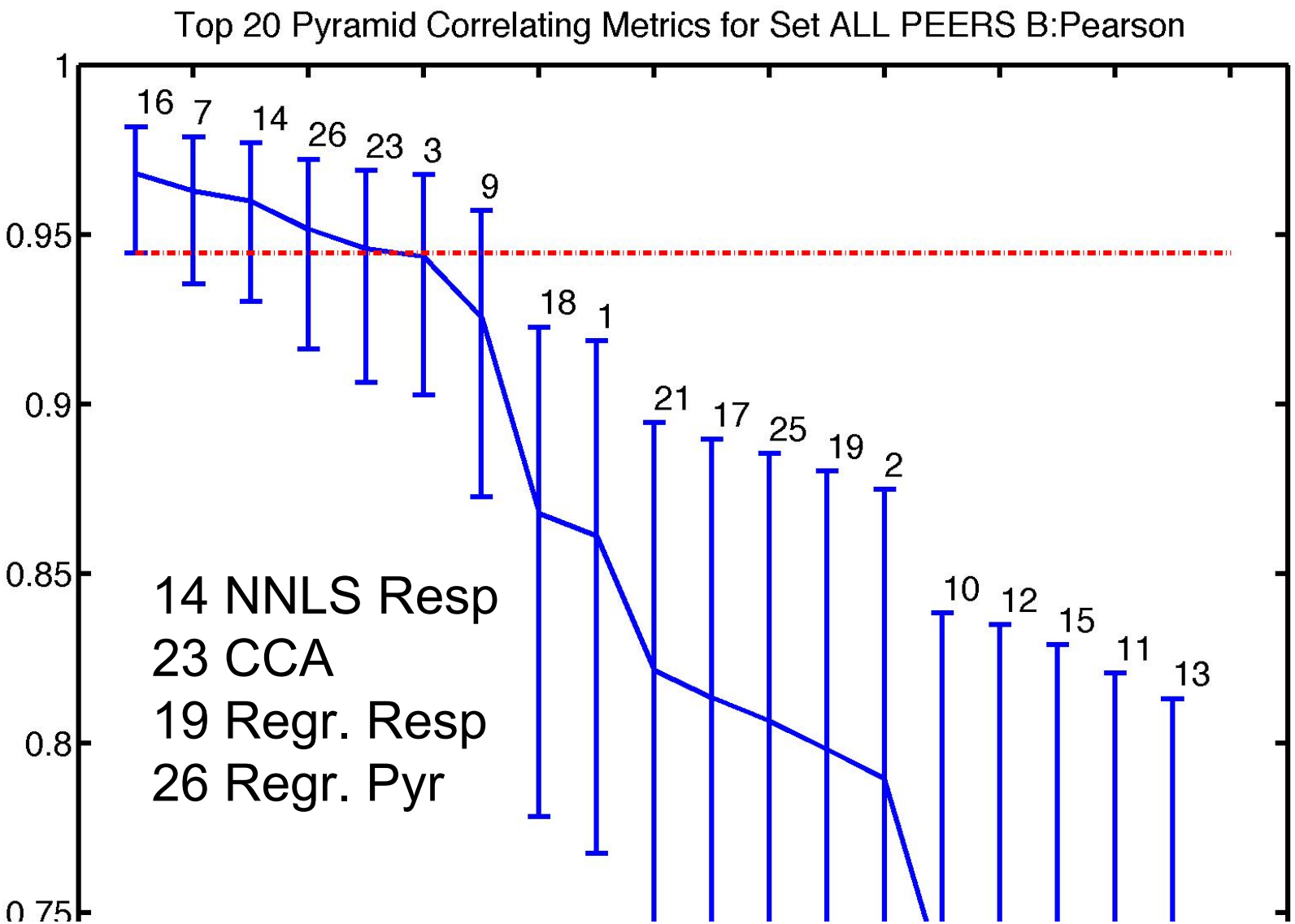
Responsiveness: Update (B)



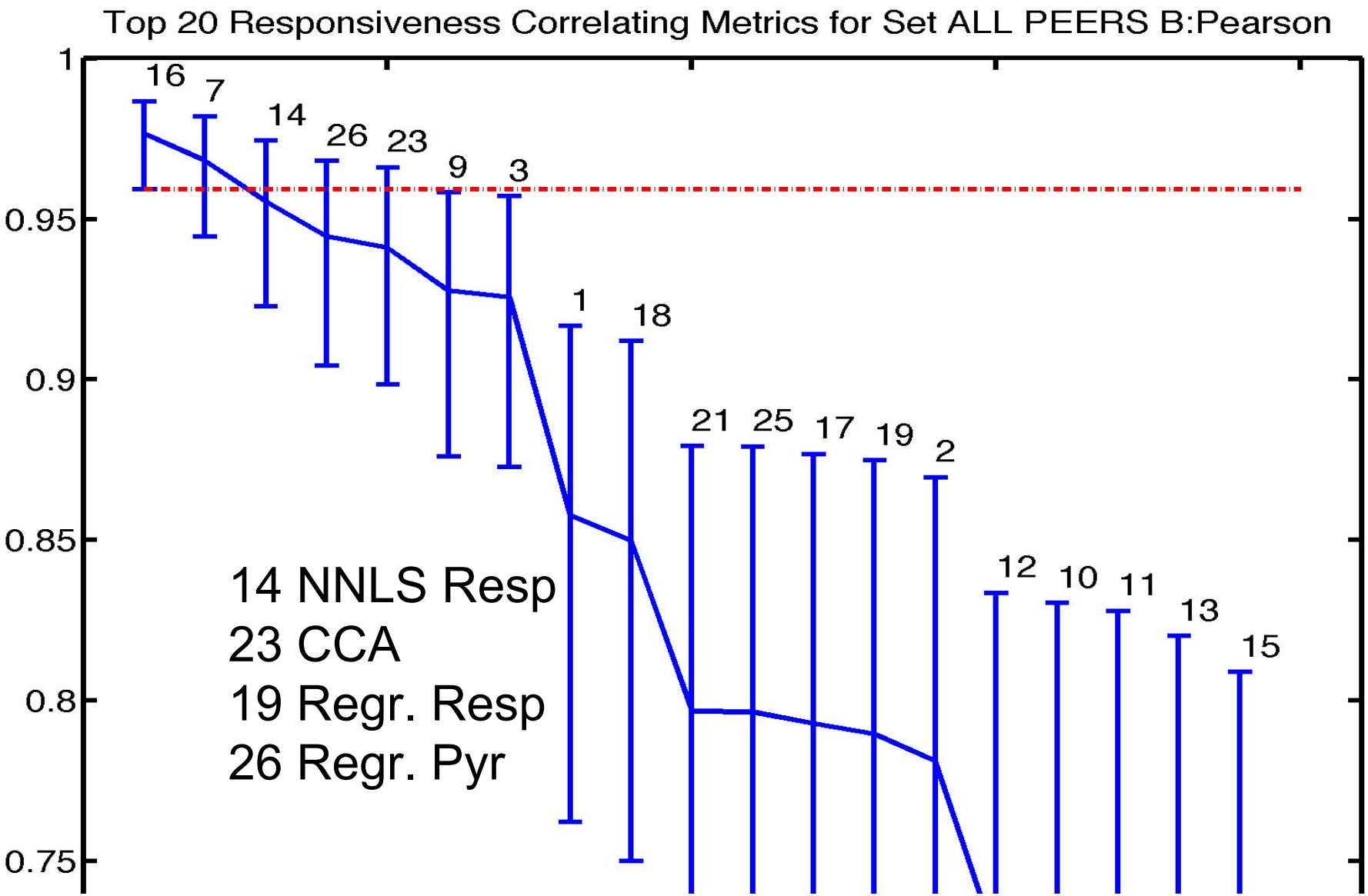
All Peers Task

- Training included human-generated summaries for TAC 2008-2009 similar to no models.
- Scoring for TAC 2010: jackknifing to compute content features.
  - Humans scored against 3 other humans.
  - Machine-generated content features are an average of scoring with 4 subsets of humans.
- Linguistic features as before.

Pyramid: Update, ALL PEERS



Responsiveness:Update, ALL PEERS



Conclusions for NO MODELS.

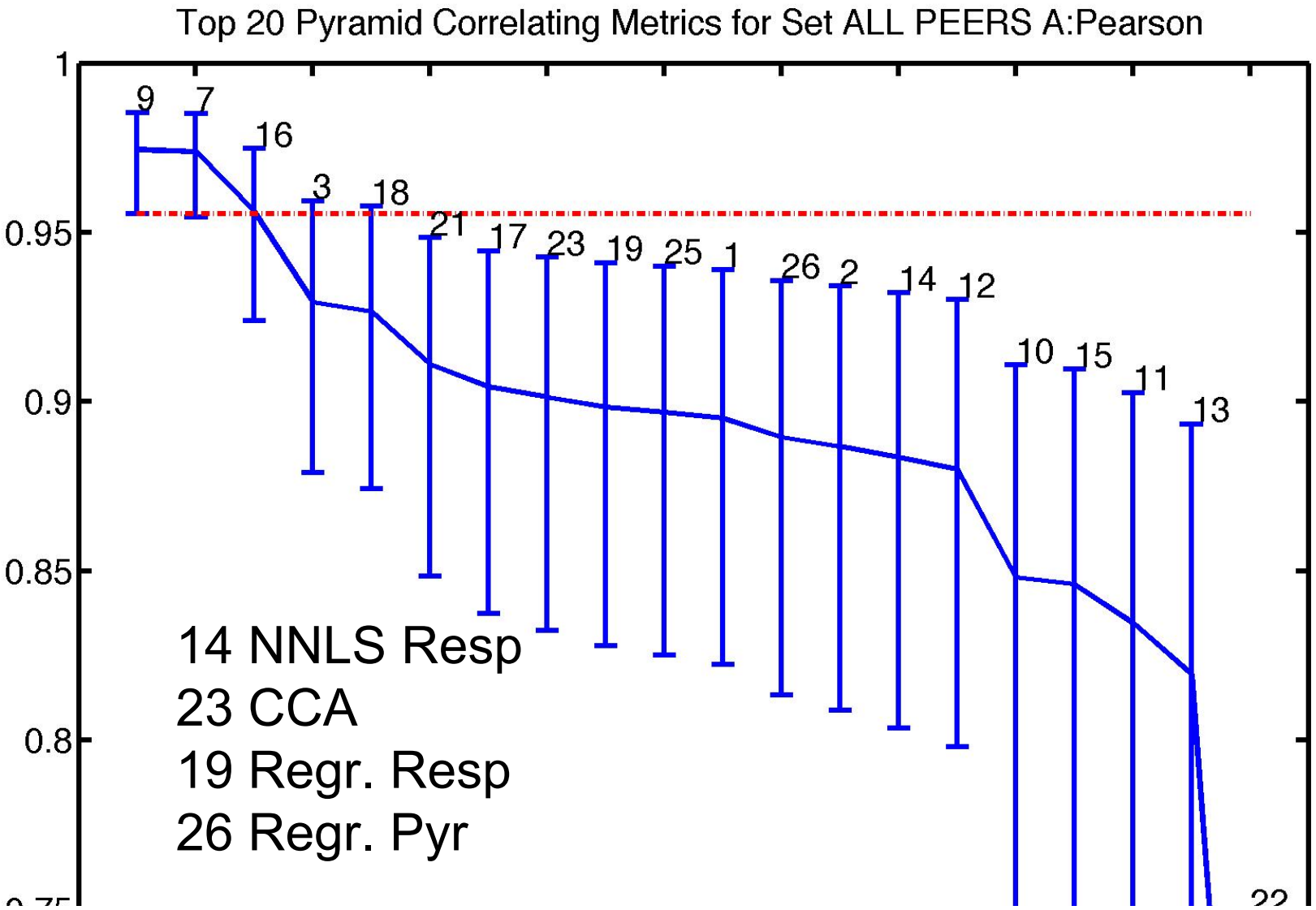
- Combining content features (ROUGE and Nouveau-ROUGE) and simple linguistic produced top metrics to predict *responsiveness*.
- A family of [wrong] CCA models are useful to build higher responsive summarization systems.
- ROUGE-2 is still strong on *pyramid*!

Conclusion for ALL PEERS and Thoughts for Future

- Nouveau-ROUGE-2 (a 2 feature model) significantly outperformed the ROUGE baselines on the update task in both responsiveness and pyramid.
- Future Work: Sharpen the linguistic features.
- Future TAC AESOP Tasks?:
  - Predicting responsiveness and linguistic SCORES.
    - Move from regression to classification.
  - Semi-automatic pyramid evaluation:Maybe an RTE Task?

Aesop's Crow and Pitcher: Persistence is Rewarded

All Peers Pyramid, Base (A)



All Peers Responsiveness.Base (A)