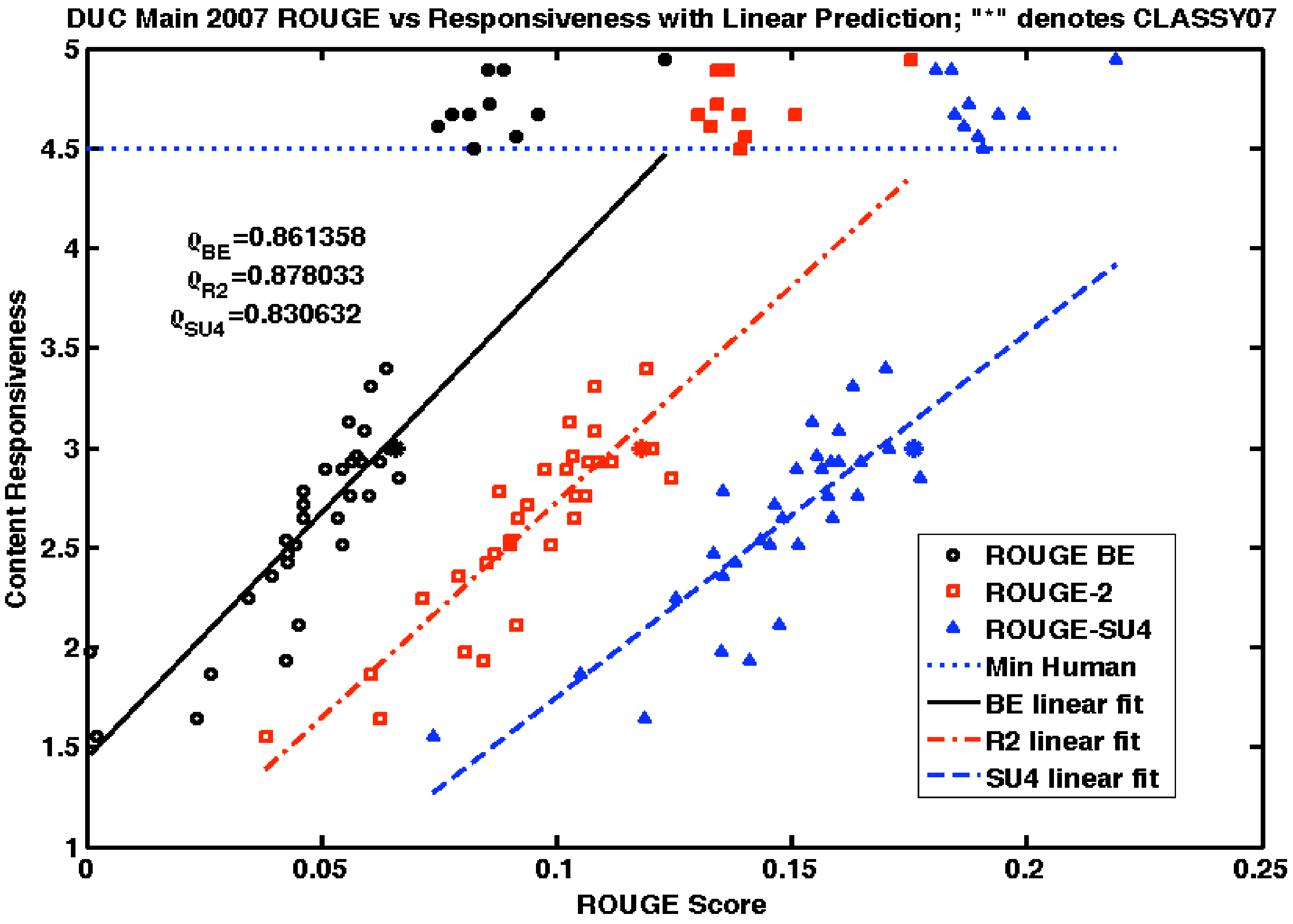
Towards Closing the Gap

John M. Conroy

IDA Center for Computing Sciences

Bowie, MD

DUC 2007: Main Task

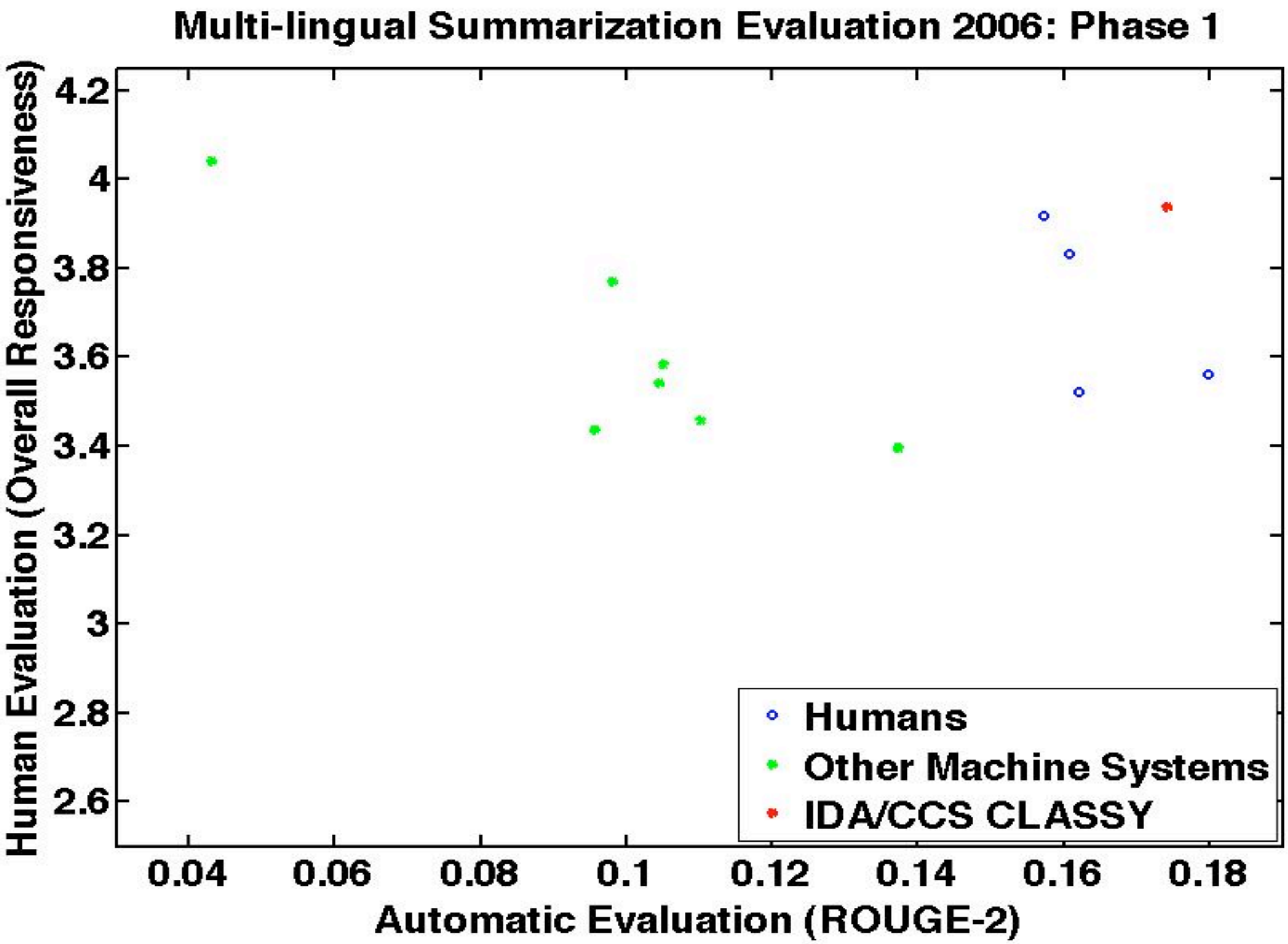


Towards Pleasing the Masses

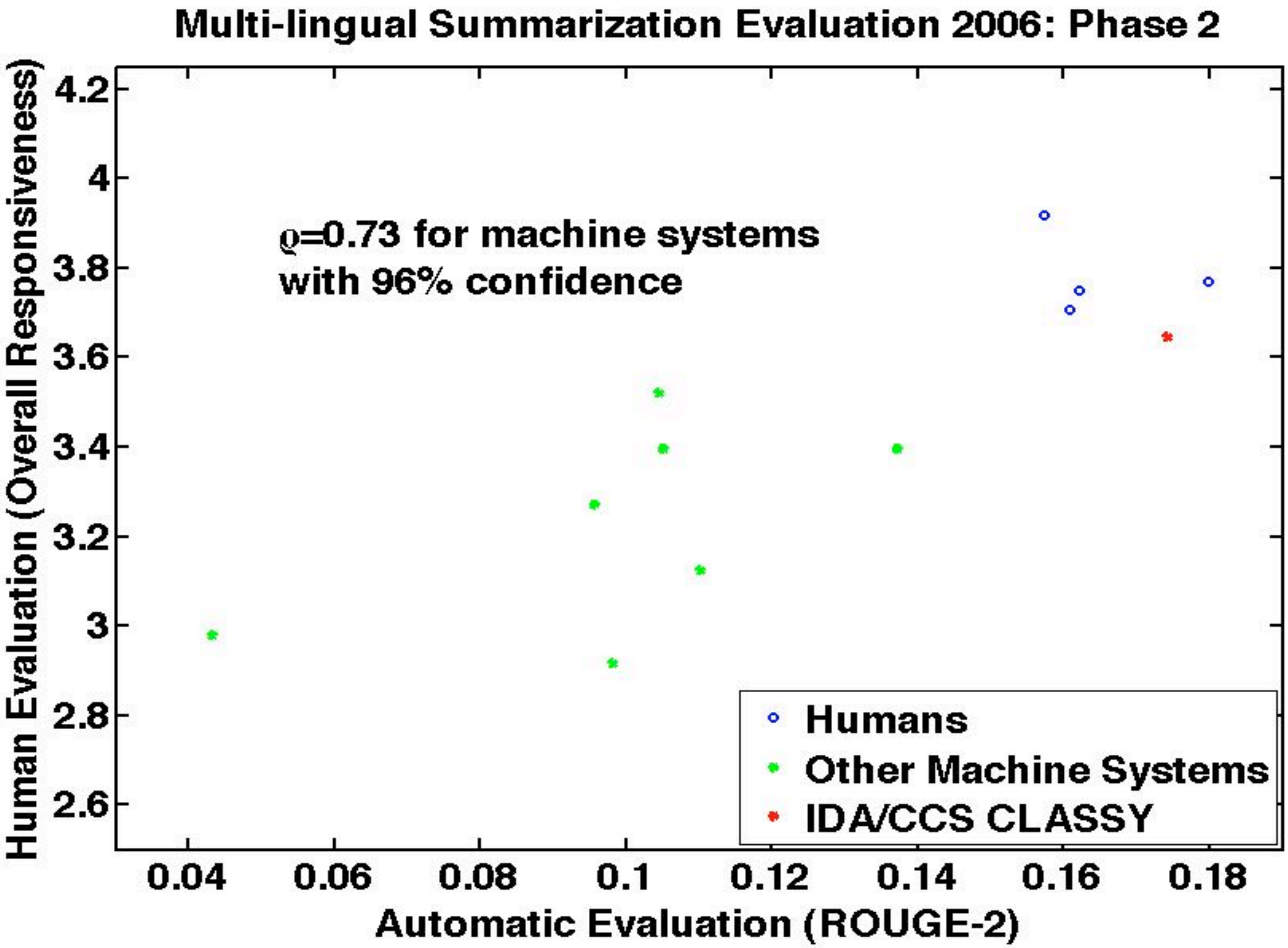
- Multi-lingual Summarization Evaluation (Lucy Vanderwende and Jade Goldstein, human evaluation setup and overseen by Jade.)
  - Assessors were NOT the summarizers.
  - Same instructions and scale as given to the DUC assessors.
  - Three Phases:

    1. Read abstracts only and assess.
       1. Very Poor, 2. Poor, 3. Barely acceptable,
       4. Good, 5. Very good
    2. Read all documents and then assess.
       (Same scale).
    3. Rank (Cluster) summaries into one of 5 equivalence classes.
       1. Unacceptable, 2. Somewhat acceptable, 3. Acceptable.
       4. Good 5. Excellent

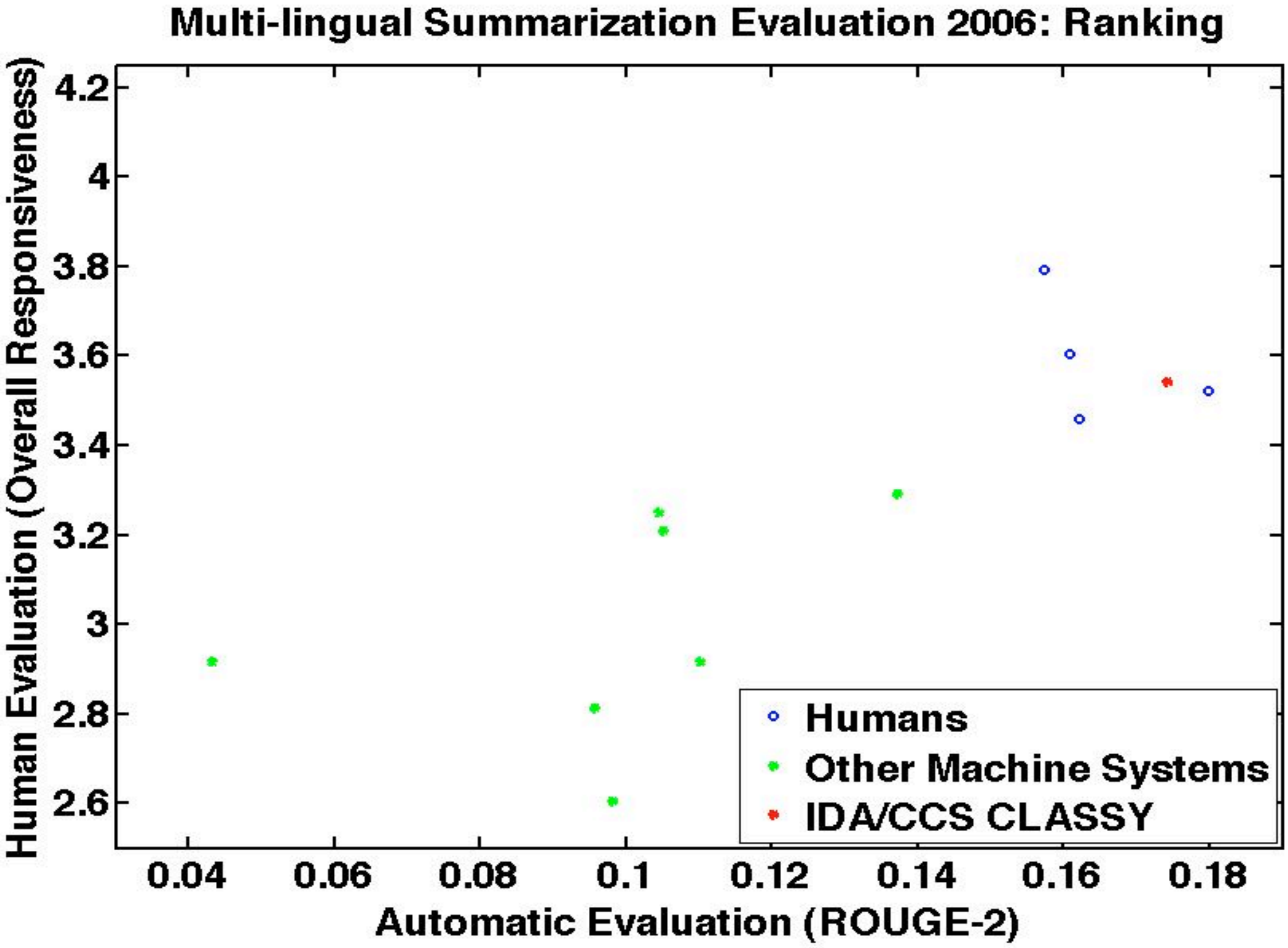
Phase 1 Results



Phase 2 Results



Ranking (Clustering) Evaluation



Some Lessons from MSE

- Assessors need to be informed on the topic to evaluate the summary.
  - Otherwise, “It’s all good!”
- Automatic Systems can please two assessors who did NOT write the summaries.
- An experiment for DUC: have 2 human evaluations from those who merely read the documents.
- Turing test vs. task based summaries.

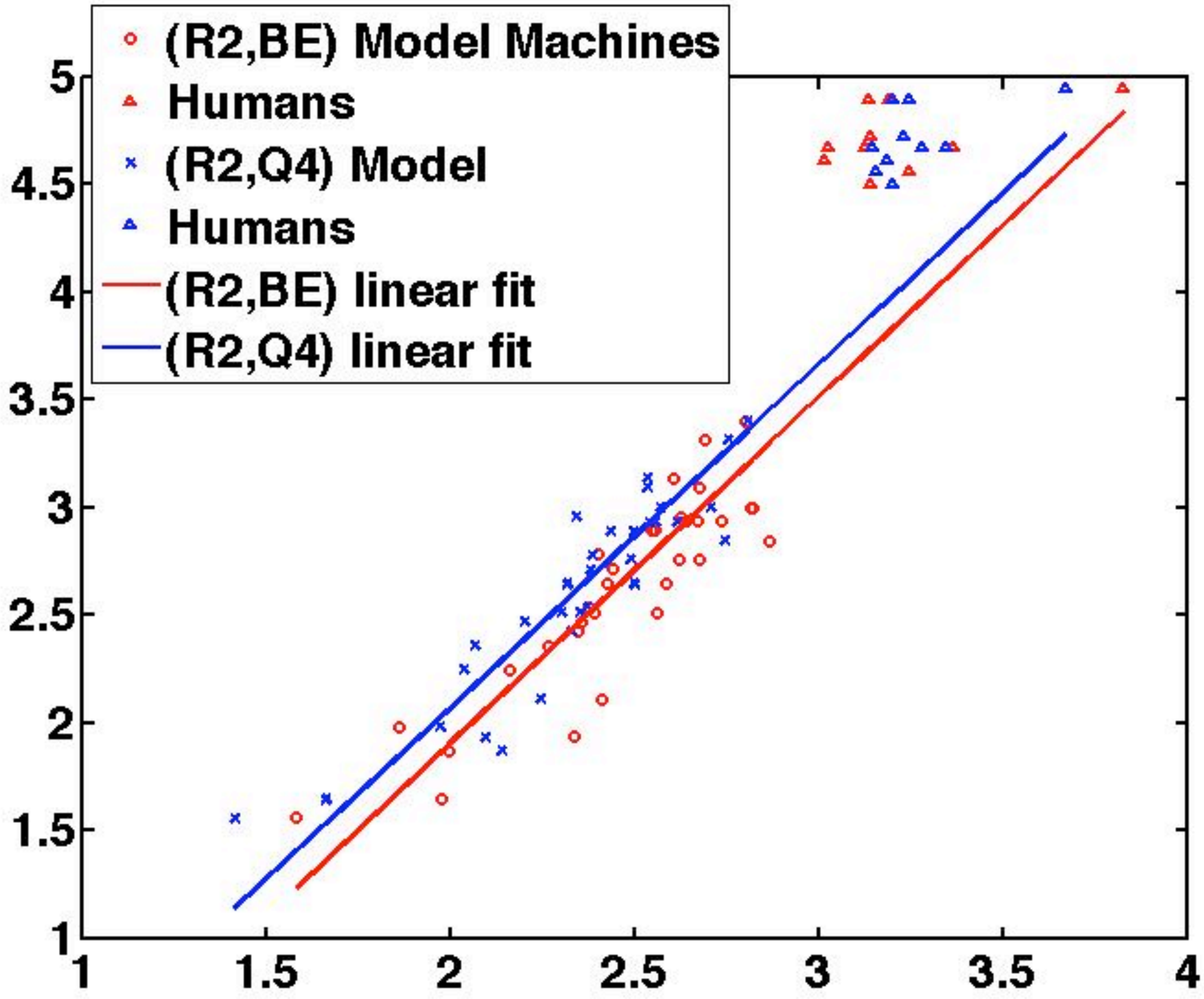
Correlating Linguistics Responsiveness

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

Modeling DUC 07 Responsiveness

- Use DUC 06 overall responsiveness, ROUGE, and linguistic questions to predict.
- Use multi-linear regression.
- Examples:
  - (ROUGE-2,Q4) 0.91.
  - (ROUGE-2,ROUGE-BE) 0.89.
- Regina Barzilay, Mirella Lapata "Modeling Local Coherence: An Entity-Based Approach", In Proc. of ACL, 2005.

Closing in on the Gap


○ (R2,BE) Model Machines
△ Humans
× (R2,Q4) Model
△ Humans
(R2,BE) linear fit
(R2,Q4) linear fit