AGGRESSIVE FILTERING WITH FASTSUM FOR UPDATE AND OPINION SUMMARIZATION

Thomson Reuters, Research & Development

Frank Schilder, Ravikumar Kondadadi, Jochen L. Leidner and Jack G. Conrad
Presenter: Jochen L. Leidner

TAC 2008, Gaithersburg, MD, USA

November 18, 2008

THOMSON REUTERS

ABOUT THOMSON REUTERS

- Leading global provider of intelligent information services to professionals
- Company brief
  - $12 bn company formed in April 2008
  - Headquarters: Thomson Reuters Tower, 3 Time Square, New York
  - Traded on New York, London, Toronto stock markets (TRI)
  - 50,000 employees in >190 countries
  - See ThomsonReuters.com

- Professional vertical markets:
  - Legal (WestLaw), News (REUTERS news), Financial markets (TR Markets), Scientific (ISI Web of Knowledge), Medical, Healthcare, Tax & Accounting

ABOUT THOMSON REUTERS R&D

- Research & Development at Thomson Reuters:
  - Group of 40+ researchers and developers
  - Chief Scientist and VP: Dr. Peter Jackson
  - Based in Minneapolis, MN and Rochester, NY, USA
  - Applied research in the following areas: information retrieval, information extraction, **summarization**, citation analysis, named entity tagging and resolution, **sentiment analysis**, data mining, record linkage, normalization and de-duplication, time series analysis, knowledge based systems, query log analysis, machine learning, personalization
  - Access to some of the largest textual, multi-medial, numerical data collections in the world

THOMSON REUTERS

Westlaw

REUTERS

calais

Powered by Thomson Reuters

OUTLINE

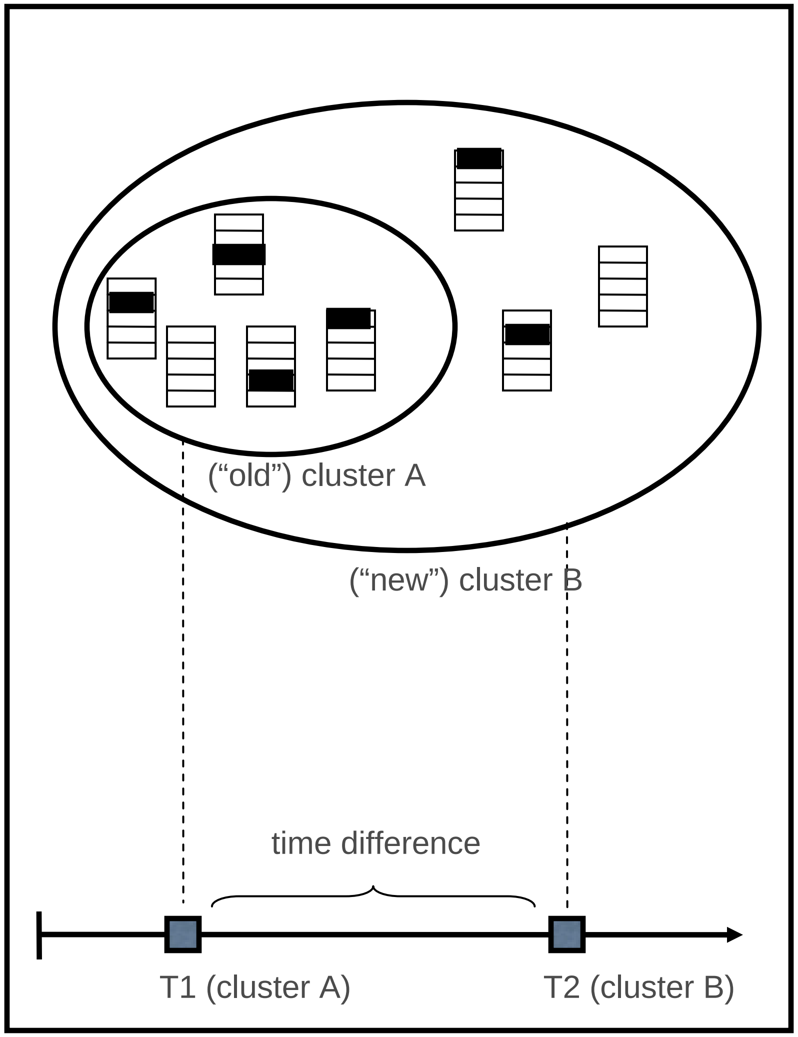
- Introduction
- FastSum system
- First sentence classifier (key innovation)
- Regression SVM
  - New features for update summarization
  - Feature selection via LARS
- Baseline
- Evaluation
- Conclusions and Future Work

INTRODUCTION

- Goals:
  - improve linguistic quality of summarization output
  - adapt FastSum:
    - to update summarization
    - develop sentiment tagger used as filter for the sentiment summarization
- Practical constraints:
  - Scalable and in near real-time
  - No complex NLP processing (e.g., parsing)
- Solution: regression SVM + feature engineering
  - Least Angle Regression (LARS)

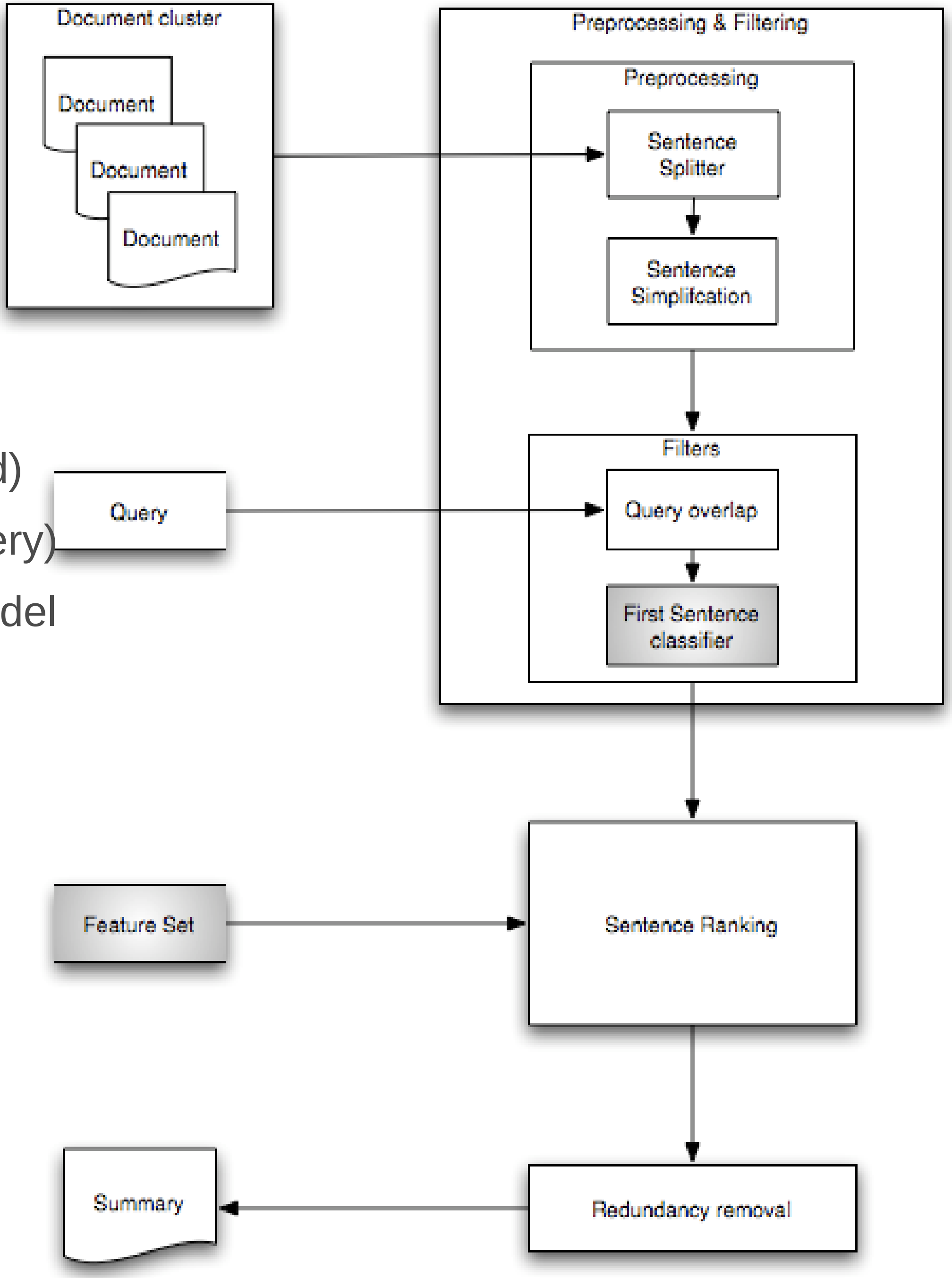
SUMMARIZATION AT TAC 2008

- I. Main Task ("Update Summarization")
  - A. (Query-Based) Multi-Document Summarization
  - B. (Query-Based) Update Summarization
- II. Sentiment Summarization Pilot Task
  - →see our poster at this conference



UPDATE SUMMARIZATION: SYSTEM DESCRIPTION

- Based on UIMA
- Processing steps:
  - Sentence splitting
  - Sentence simplification (lexicon-based)
  - Filter ((in-)exact word overlap with query)
  - Sentence ranking via a regression model
  - Redundancy removal (QR decomposition)
- Regression SVM (Li et al., 2007)
  - Define summary-worthy sentence by word overlap with model summaries
  - Create simple, efficiently to compute features
  - Trained on DUC 2007 data



LINGUISTIC QUALITY

- Improvement of linguistic quality by
  - sentence simplification (already done before)
  - name simplification
    - keeping track of names (*George W. Bush*), generate abbreviated name (*Bush*)
    - mentioning long name first, abbreviated name later
  - first sentence classifier
    - first sentences often can be seen as a very concise summary of the entire article
    - first sentence-like sentences reduce dangling references (rhetorical, pronouns etc.)

FIRST SENTENCE CLASSIFIER

- Key innovation
- Classifies whether or not a sentence *s* is similar in nature to typical first sentences of articles
- Motivation: improve linguistic quality by avoiding dangling references (e.g. *therefore, he, after that, …*):

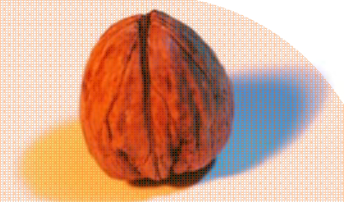
  ~~Therefore~~*, Colin Powell endorsed Barack Obama.*

- Features: capitalized words, pronouns, definite articles, words, connectors, quotes
- Trained on first and non-first sentences of randomly chosen 50k documents from AQUAINT-2

FASTSUM: EXAMPLE SUMMARY

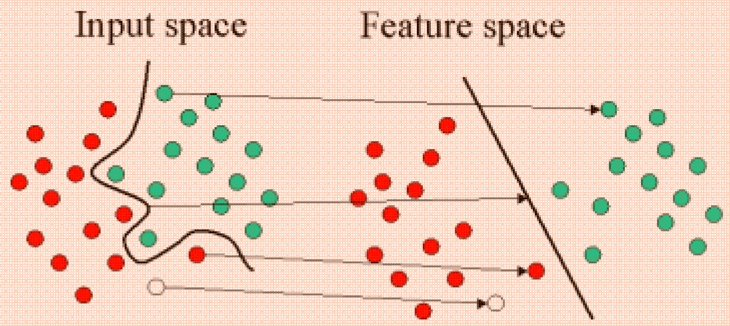
- QUERY:
  Describe India's space program efforts and cooperative activities with other nations in space exploration.

- SUMMARY:
  The United States, the European Space Agency, China, Japan and India are all planning lunar missions during the next decade. The U.S. space agency NASA is in talks with its Indian counterpart on whether to take part in New Delhi's first unmanned moon mission set for 2007. The European Space Agency and National Aeronautics and Space Administration's X-ray and laser equipment will ride piggyback on India's Chandrayaan-1. The space agencies of India and France signed an agreement to cooperate in launching a satellite in four years that will help make climate predictions more accurate.

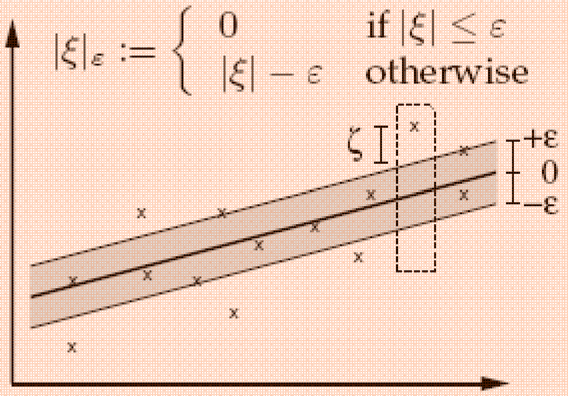


REGRESSION SVM IN A NUTSHELL

- Support Vector Machines (Vapnik & Lerner, 1963):
  - Map to higher-dimensional feature space in order to achieve linear separability
  - Use maximal margin (best-separating decision boundary in hyper-plane)

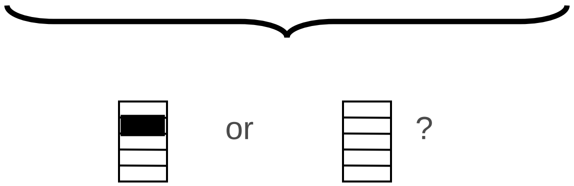


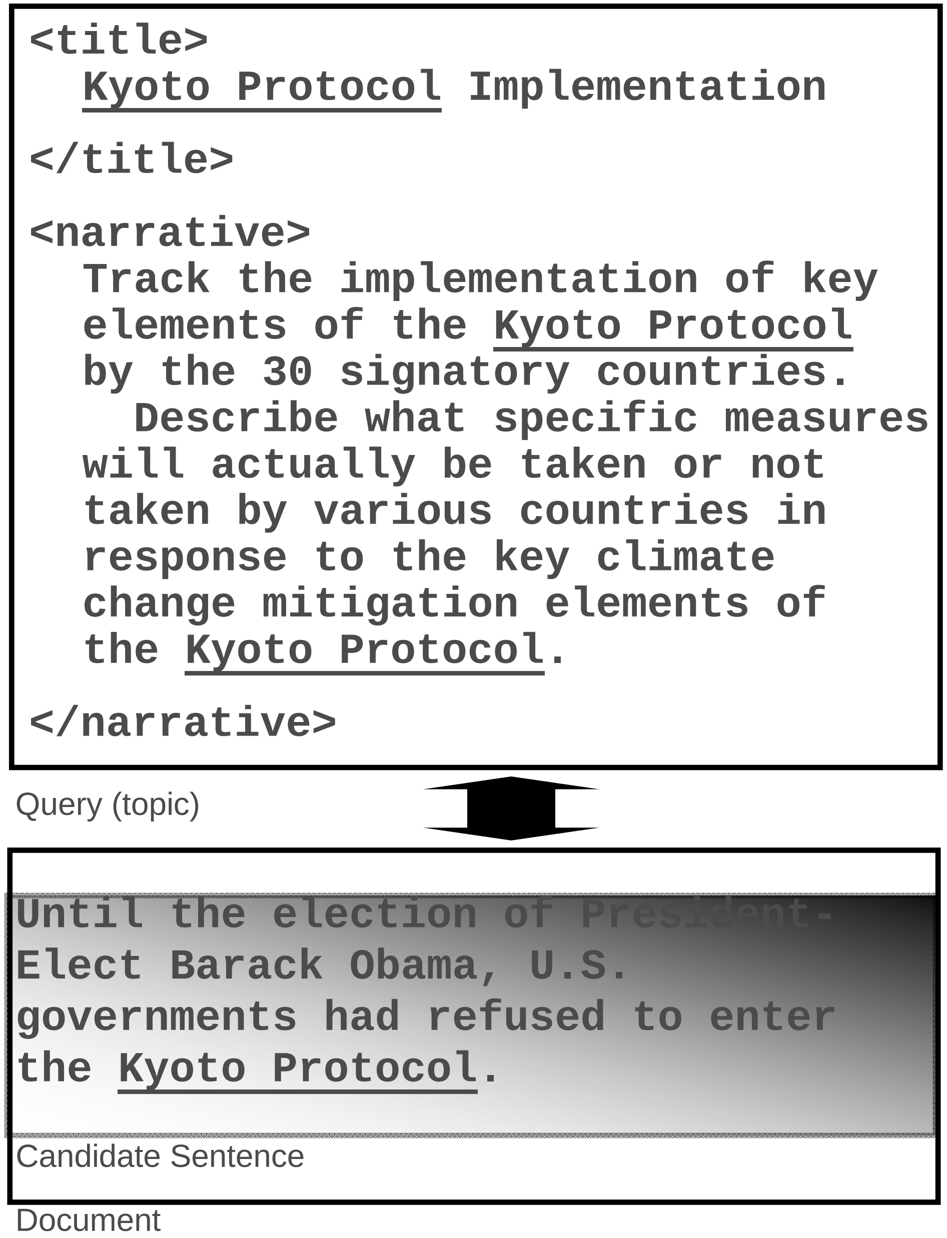
- Regression SVM (Vapnik et al. 1996; Schölkopf & Smola, 1998)
  - Apply SVMs to non-Boolean objective functions
  - Implemented using *SVM^{light}* package (Joachims, 1999)


$|\xi|_\varepsilon := \begin{cases} 0 & \text{if } |\xi| \le \varepsilon \\ |\xi| - \varepsilon & \text{otherwise} \end{cases}$
ζ
+ε
0
−ε
x

FEATURES (SCHILDER & KONDADADI, 2008)

1. **Topic title frequency**
2. Topic narrative frequency
3. Content word frequency
4. **Document frequency**
5. Headline frequency
6. Sentence length (binary/integer)
7. Sentence position (binary/integer)





```
<title>
  Kyoto Protocol Implementation
</title>
<narrative>
  Track the implementation of key
  elements of the Kyoto Protocol
  by the 30 signatory countries.
    Describe what specific measures
  will actually be taken or not
  taken by various countries in
  response to the key climate
  change mitigation elements of
  the Kyoto Protocol.
</narrative>
```

Query (topic)

```
Until the election of President-
Elect Barack Obama, U.S.
governments had refused to enter
the Kyoto Protocol.
```

Candidate Sentence

Document

NEW FEATURES (1/2)

1. “Old” (= Cluster A, T1) Content Word Frequency
   – relative content word frequency $p_c(t_i)$ of all *old* content words $t_{\{1..|s|\}}$ occurring in a sentence *s*

2. Old Document Frequency
   – relative document frequency $p_d(t_i)$ of all *old* content words $t_{\{1..|s|\}}$ occurring in a sentence *s*

3. Old Entities
   – number of named entities in the sentence that already occurred in the old cluster

4. “New” (= Cluster B, T2) Entities
   – number of new named entities in the sentence not already mentioned in the old cluster

NEW FEATURES (2/2)

1. Old/New Entity Ratio
   - ratio of number of unseen named entities in the sentence to number of named entities in the sentence that were already seen
2. New Words
   - number of new content words in the sentence not already mentioned in the old cluster
3. Old Words
   - number of content words that already occurred in the old cluster
4. Old/New word ratio
   - the ratio of the number of old and new word

LARS (LEAST ANGLE REGRESSION)

- Efron et al., 2004
- model selection algorithm to find a minimal set of features
- Best combination of features can be algorithmically determined
- Features that are most correlated with the response added to model incrementally
- Coefficient of the feature is set in the direction of the sign of the feature's correlation with the response
- R package: http://cran.r-project.org/web/packages/lars/index.html
- Reduced feature set from 12 to 5

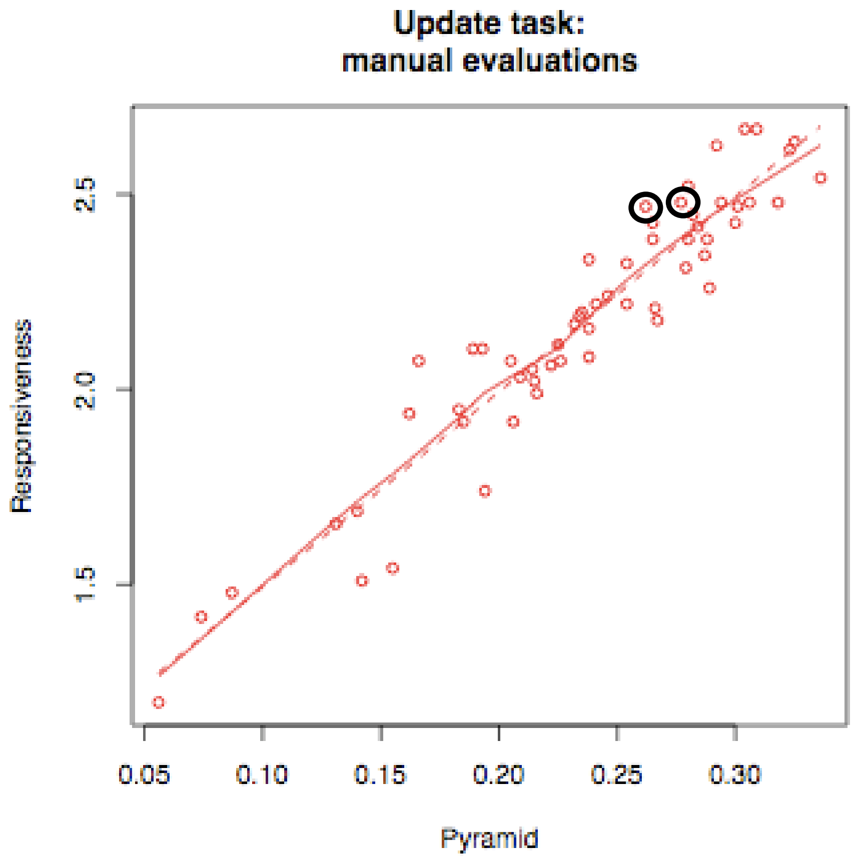
BASELINE

- Simple baseline:
  - extract first sentences from each document in the cluster
  - sort sentences according to the document's timestamp
  - eliminate redundancy by adding only sentences with cosine similarity <0.7 for candidate sentence and summary so far
- Our own proposed baseline system
- Ranked 13th for ROUGE-2 compared to 28th/29th rank for our other two systems

MANUAL EVALUATION

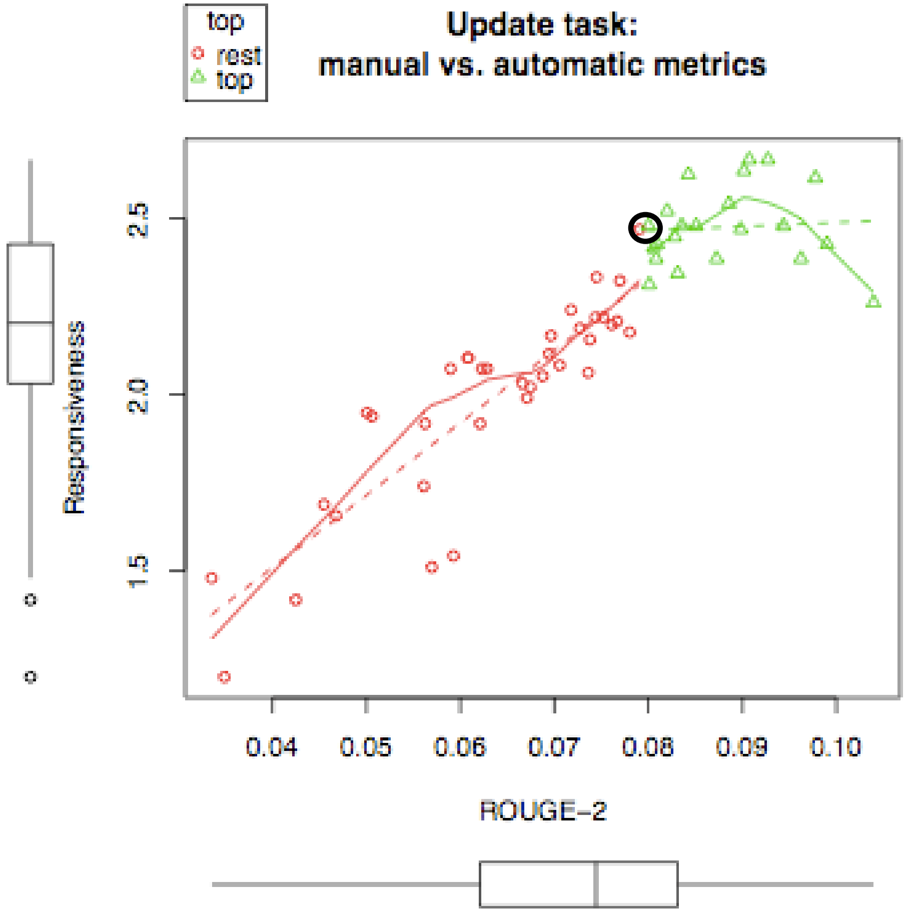
- NIST assessors scored summaries from 1 (very poor) to 5 (very good) by
  - linguistic quality (taking into account non-redundancy, focus, structure, and coherence)
  - responsiveness (taking into account the information need)

- Received
  - **Highest score (tied 1st/58) for first sub-task**
  - very high scores for linguistic quality (→first sentence classifier) (4th/58)
  - very high scores (7th/58) for responsiveness
  - average scores for pyramid score (despite correlation with human responsiveness)



AUTOMATIC EVALUATION

- Automatic ROUGE scores average compared to competitors
- 28th/29th out of 71 participant systems
- ROUGE-2/Responsiveness for top-22 systems not correlated (Pearson)



THOMSON REUTERS

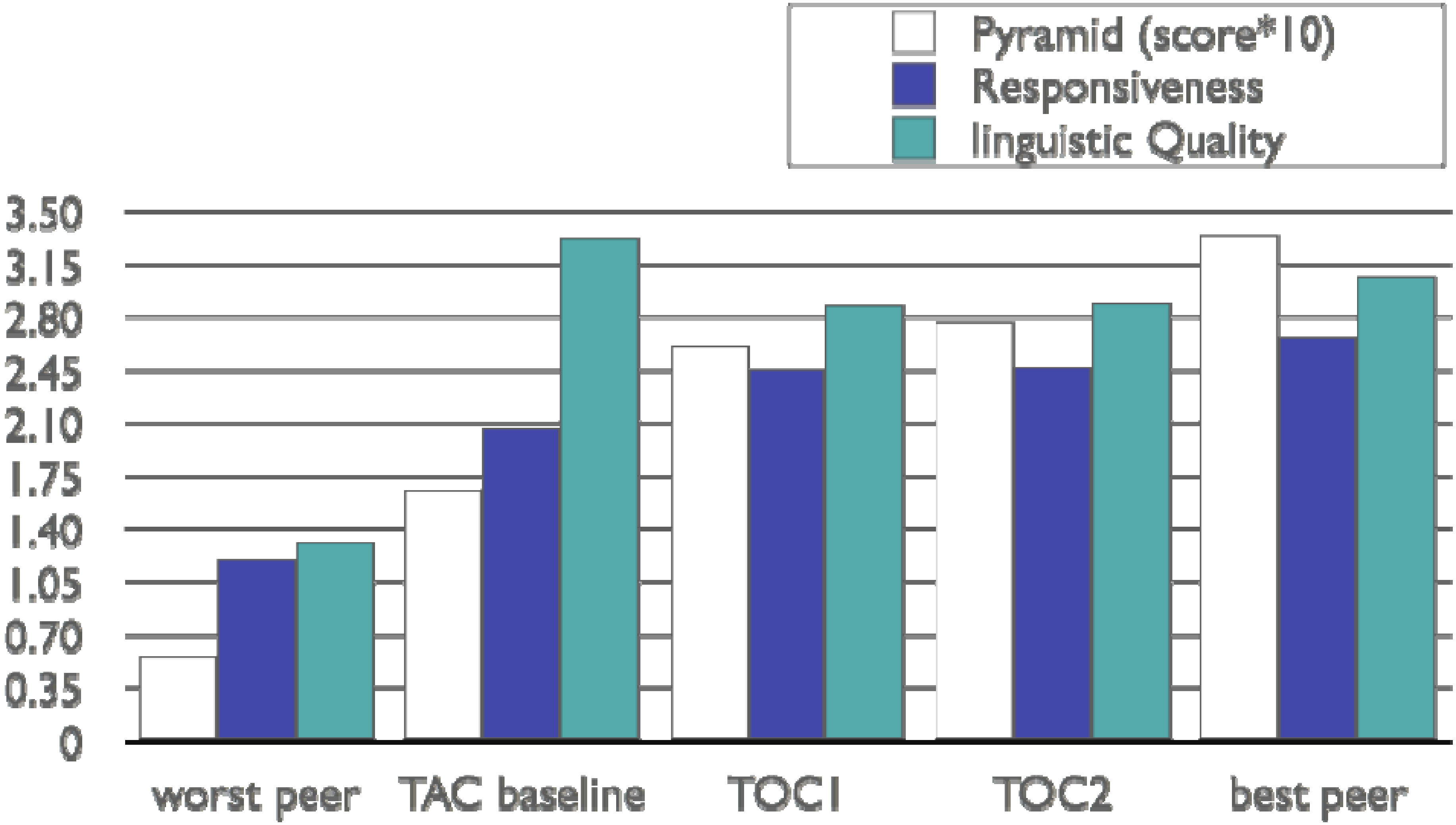
CONCLUSIONS & FUTURE WORK

- First sentence classifier improved linguistic quality (best run ranked 4th for overall linguistic quality)
- Update summaries generated via a regression SVM using features such as number of old/new entities
- Optimal number of features determined LARS feature selection algorithm (similar performance compared to full feature set)
- Run with all features ranked 7th for responsiveness, run with the optimized feature set ranked 8th.
- We proposed a simple baseline for the update task that received high ROUGE scores.
- A sentiment summarization system was built based on FastSum for Sentiment Pilot task (→**poster session**)
- Future work:
  - find better features for updates
  - better automatic evaluation?
  - incorporate *credibility* (Conrad, Leidner & Schilder, 2008) in the sentence selection algorithm (next TAC pilot task?)

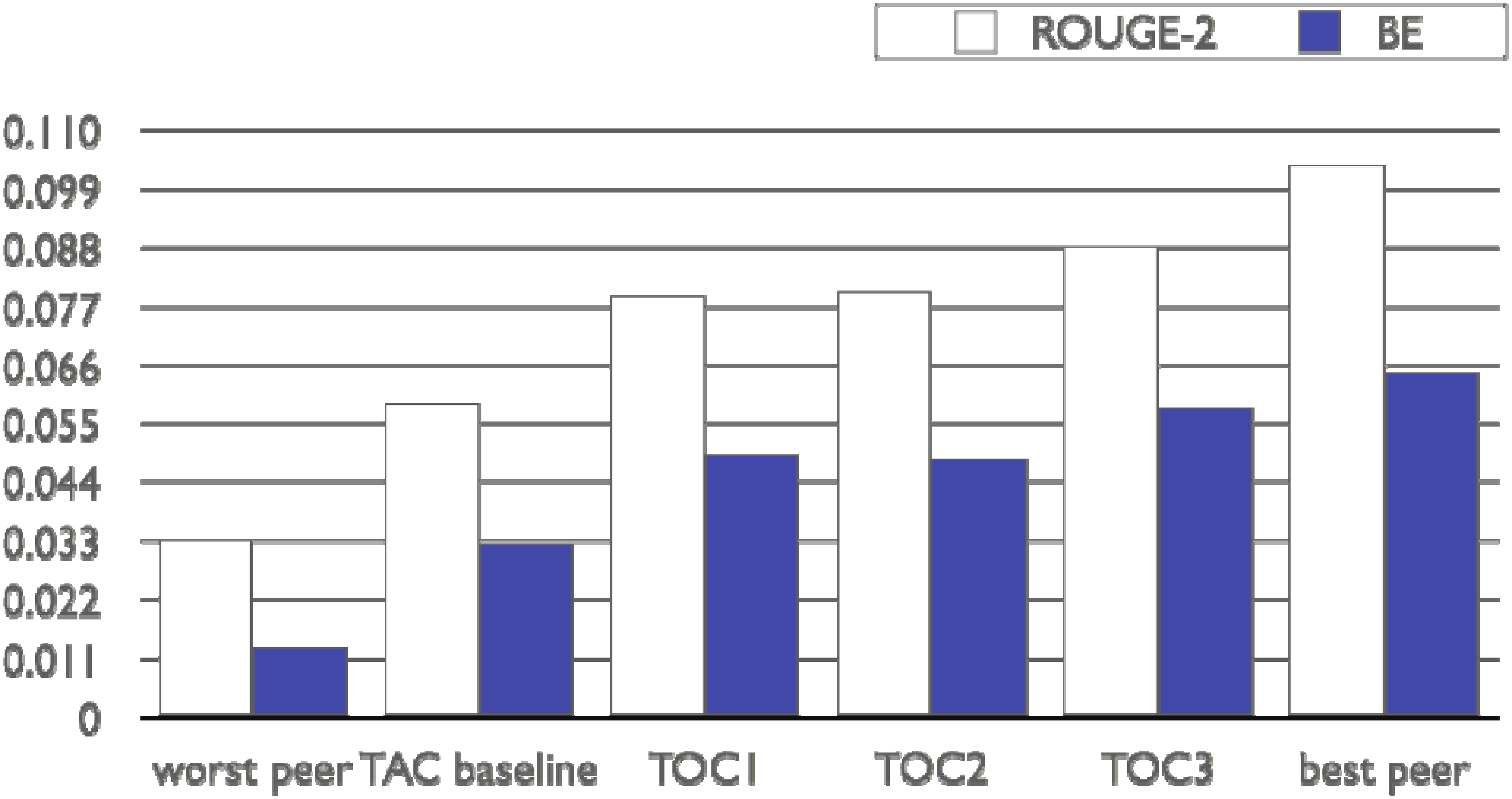
QUESTIONS?

BACKUP SLIDES

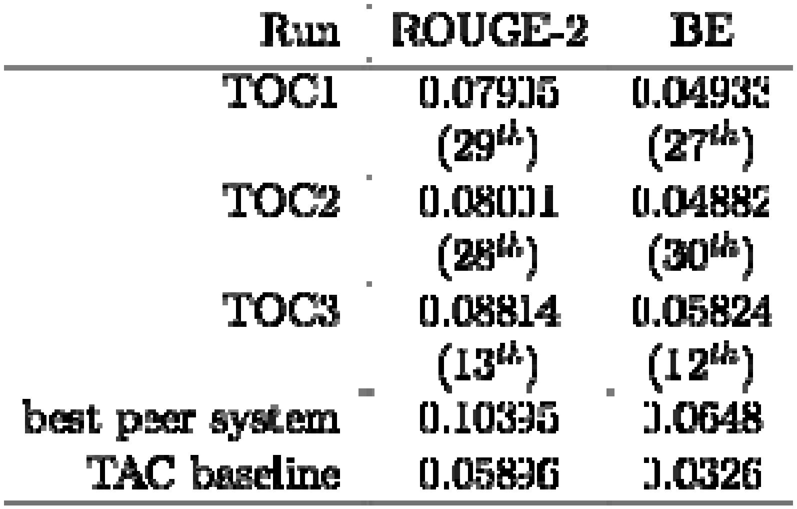
TAC 2008 RESULTS



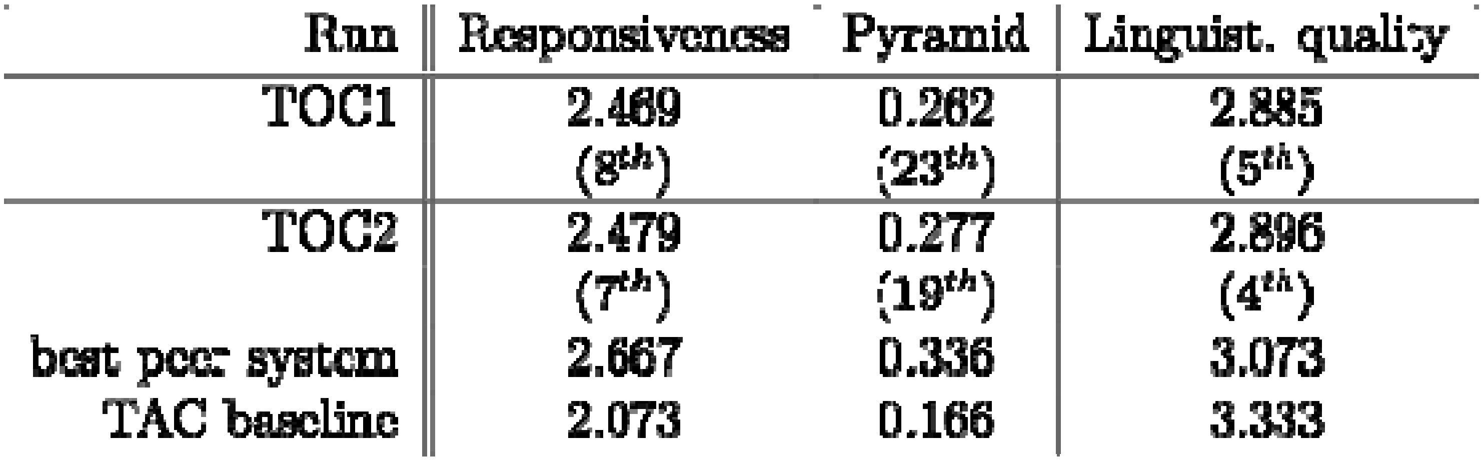
TAC 2008 RESULTS



TAC 2008 RESULTS

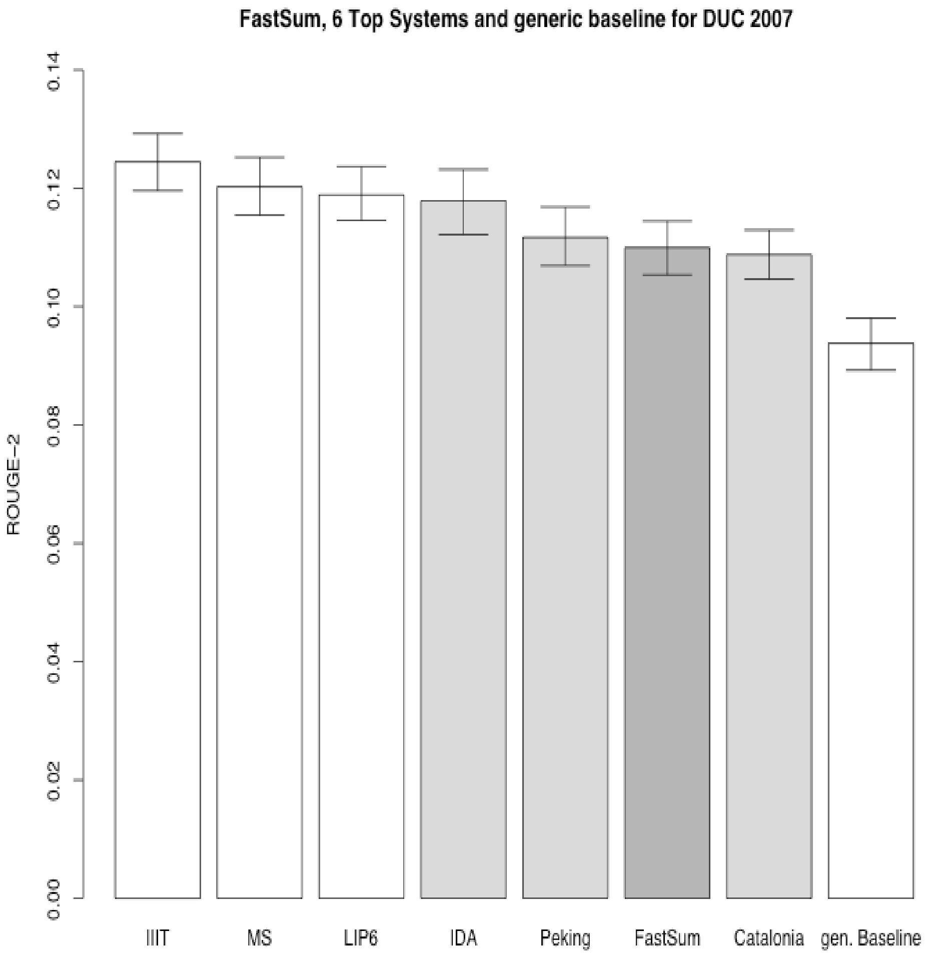


Run	ROUGE-2	BE
TOC1	0.07935 (29th)	0.04933 (27th)
TOC2	0.08031 (28th)	0.04882 (30th)
TOC3	0.08814 (13th)	0.05824 (12th)
best peer system	0.10395	0.0648
TAC baseline	0.05896	0.0326



Run	Responsiveness	Pyramid	Linguist. quality
TOC1	2.469 (8th)	0.262 (23th)	2.885 (5th)
TOC2	2.479 (7th)	0.277 (19th)	2.896 (4th)
best peer system	2.667	0.336	3.073
TAC baseline	2.073	0.166	3.333

REVIEW OF PAST RESULTS

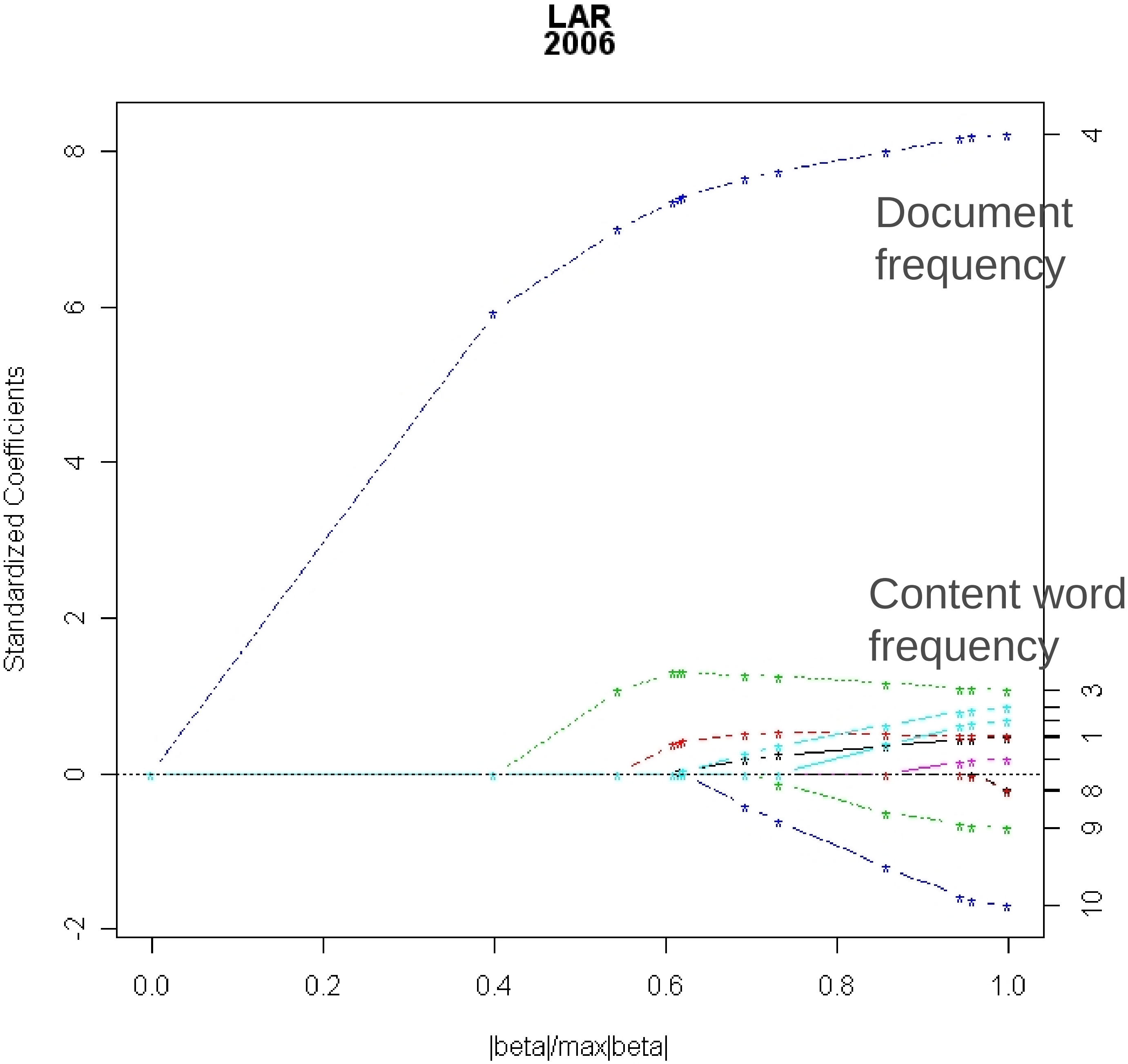


- DUC 2007 (post-hoc)
  - Rank 6
  - Rouge-2: 0.11095

- DUC 2006 (post-hoc)
  - Rank 2
  - Rouge-2: 0.0925

LARS

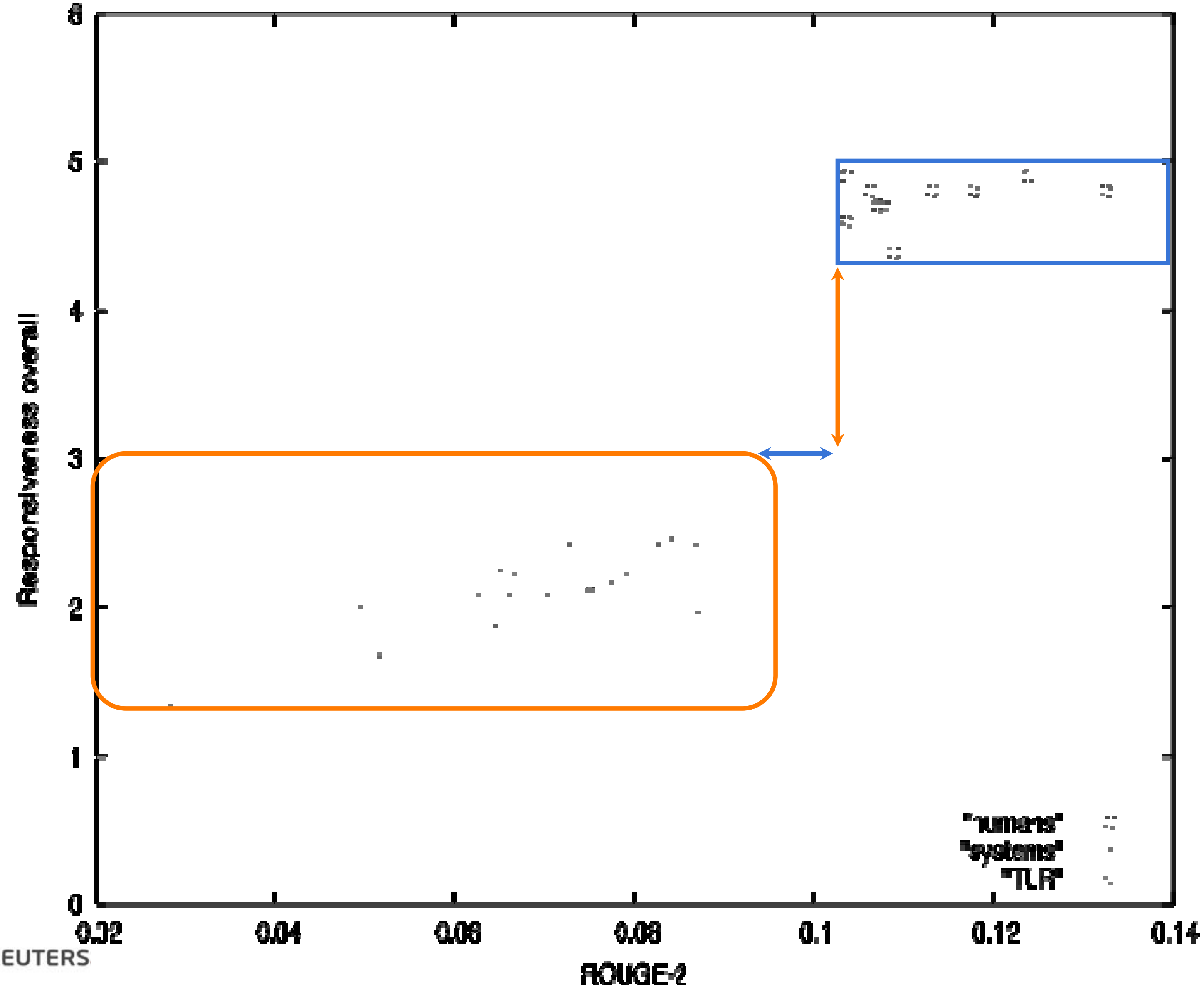
- Features are plotted along the x-axis
- The corresponding coefficients are shown on y-axis.
- The earlier a feature appears on the x-axis, the more important it is.



Document frequency

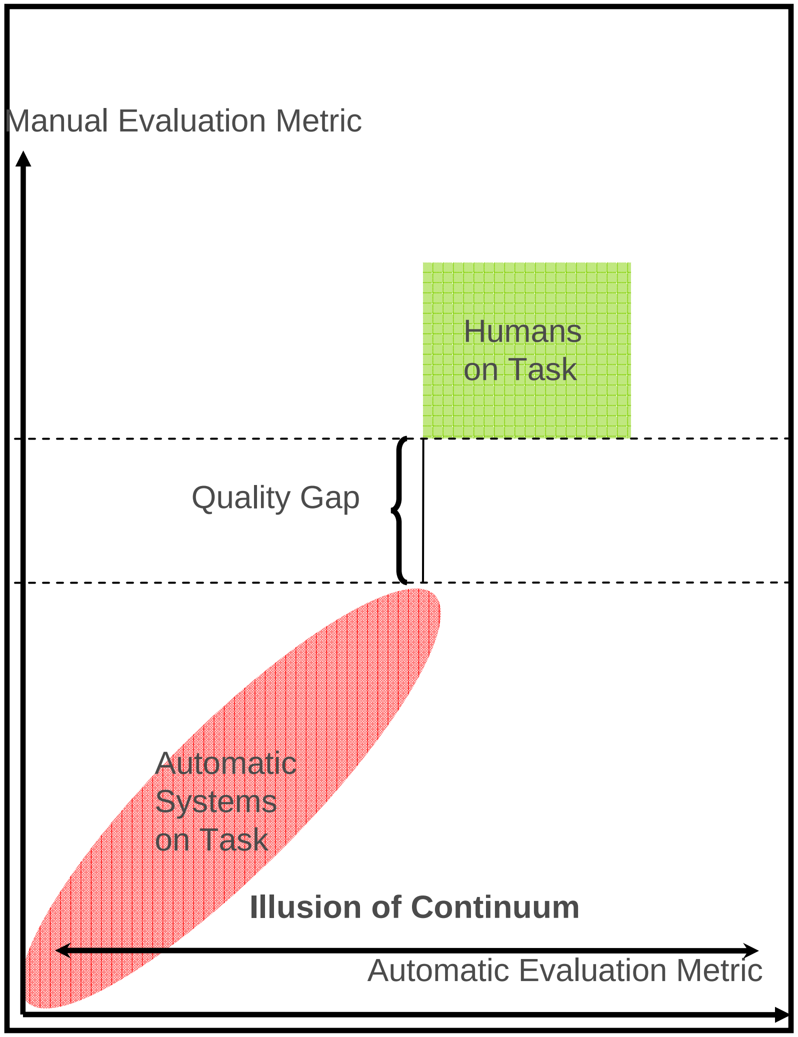
Content word frequency

DUC 2006 RESULTS



DISCUSSION: AUTOMATIC EVALUATION

- Correlation between manual and automatic quality measures reveals **quality gap**
- Automatic metrics suggest illusion of a continuum
- Manual metric exposes wide gap between “bad” and “good” raters/systems
- Ceiling appears to have been reached
- First described by Schilder et al. (2006)
- → ***Can automatic evaluation metrics only assess bad systems?***



REFERENCES

- Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani (2004), “Least angle regression**”** *Ann. Statist.* **32***(2)*: 407-499.
- Schölkopf & Smola (1998) *A Tutorial on Support Vector Regression.* NeuroCOLT Technical Report.