# Guiding CLASSY Toward More Responsive Summaries

***Judith D. Schlesinger***

***IDA Center for Computing Sciences, USA***

***Joint work with***

***John M. Conroy, IDA Center for Computing Sciences***

***Dianne P. O'Leary, U. of Maryland***

***Jeff Kubina, Dept. of Defense***

# Overview

- CLASSY
  - What's the same
  - What's new
- How we did
- Future efforts

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

- Data Preparation
- Linguistic Preprocessing
- Sentence Scoring
- Redundancy Removal
- Sentence Ordering

# Data Preparation

- NEW: Create training data
  - Use TAC 2008 and 2009 data

# Data Preparation

- NEW: Create training data
  - Use TAC 2008 and 2009 data
- Query term identification
  - Topic title (+ description when available)
  - NEW: category terms and aspect terms

# Data Preparation

- NEW: Create training data
  - Use TAC 2008 and 2009 data
- Query term identification
  - Topic title (+ description when available)
  - NEW: category terms and aspect terms
- Submission 16: topic title + category terms
- Submission 13: as before + type aspect terms
- ?? How to use remaining aspects?

# Example

accident => [“calamity”, “casualty”, “collision”, “crash”, …]

(type) [“explosion”, “cave in”, “collapse”, “train”, …]

(when) [months, days, yesterday, today, tomorrow, …]

(where) [“location”, “position”, “place”, “nearby”, …]

(why) [“broken”, “faulty”, “careless”, “fatigue”, …]

(affected) [“death toll”, “decease”, “wound”, “survivor”, ...]

(damages) [“propery loss”, “ruin”, “goods”, “economic”, …]

# Linguistic Preprocessing

- Classify “sentences”
  - -1: don’t use
  - 0: statistics only
  - 1: statistics and selection
- Sentence splitter: FASST-E
  - < .01% (known) error rate
  - 1000+ sentences/second

# Linguistic Preprocessing (cont.)

- Tokenize and trim
  - Boilerplate removal
  - Phrase/clause trimming
  - NEW: quotation mark matching
- No POS-tagging or parsing
  - Pattern-matching rules only

# Trimming Patterns

- Eliminations
  - Gerund phrases
  - Relative clause appositives
  - Attributions
  - Lead adverbs and phrases
    - 100+ “patterns” ranging from 1 to 300+ word variations per pattern ➔ ~2000 phrases
    - At the same time, By the way, …

# Eliminations (cont.)

- Medial adverbs/phrases
  - ~25 “patterns” ranging from 1 to 80+ word variations per pattern ➔400+ phrases
  - too, however, in fact, for example
- Age references
  - , 32,  , age 29,
- Parenthesized phrases
- Dashed phrases

# Sentence Scoring

- An Approximate Oracle
  - Use signature terms
  - Query terms
    - From document set titles (and descriptions)
    - NEW: aspect descriptions

# Approximate Oracle

- An oracle depends on knowledge of human abstracts
- Need method that doesn't rely on having abstracts
- Use signature terms as “samples” from idealized human summaries

# Signature Terms

- Term: stemmed (lemmatized), space-delimited string of alphabetic characters
  - all text is made lower case
  - non-text characters and stop words are removed
- Signature term: terms that occur more often than expected
  - Based on a 2×2 contingency table of relevance counts

# Approximate Oracle (cont.)

- Approximate the Oracle by $P(t|\tau)$, a sum of 3 distributions:
  1) $s_t(\tau)$ = 1 if term *t* is signature term for topic $\tau$
     = 0 otherwise
  2) $q_t(\tau)$ = 1 if term *t* is query term for topic $\tau$
     = 0 otherwise
  3) NEW: $p_t(\tau)$ = maximum likelihood estimate of the probability that term *t* occurs in a sentence with one or more signature/query terms in $\tau$

# Approximate Oracle (cont.)

- NEW: weighted by (0.3, 0.3, 0.4)
  - Based on training data subset of TAC 2008-2009
- Sentence score = sum of $P(t|\tau)$ taken over all its terms divided by its length
- NEW: Bias noted against first sentence
  - score adjusted based on training data

# Nouveau Projection: Update

- Create the term-sentence matrix for the base summary
- Create the term-sentence matrix of new information
  - NEW: Non-negative matrix factorization with weighting determined from Nouveau-ROUGE training

# Redundancy Removal

- Approximate oracle selects candidate sentences
- LSI improves the score by giving “partial credit” for loosely related terms not actually in the same sentence
  - Moves from a term-based to an idea-based measure by projecting sentences onto a subspace of ideas

# Redundancy Removal (cont.)

- (Non-negative) pivoted-QR selects sentences that provide distinct information
- NEW: Integer programming chooses sentences that best utilize the 100-word space for summaries
  - Improved average words per summary from 94 (2009) to 98 (2010)

# Sentence Ordering

- Models output as Traveling Salesman Problem to find shortest path among sentences
  - Term overlap used to measure similarity
  - Monte Carlo method used to approximate solution of NP-hard problem

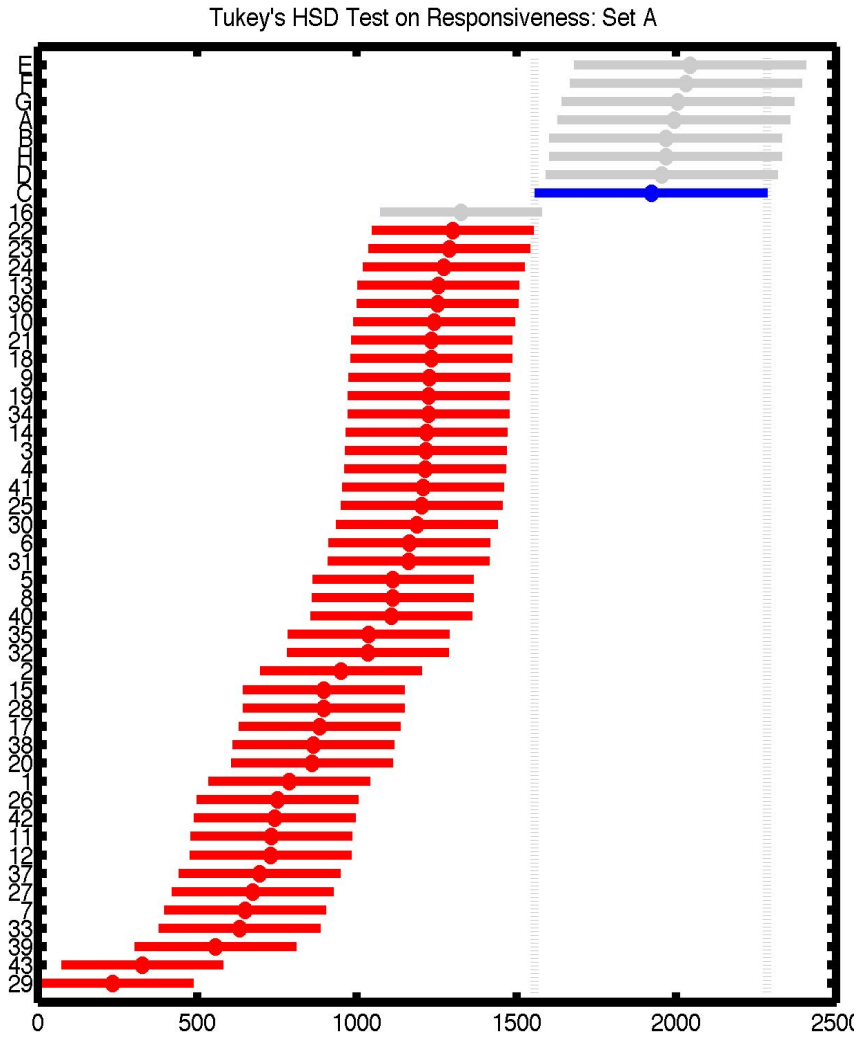
# Results

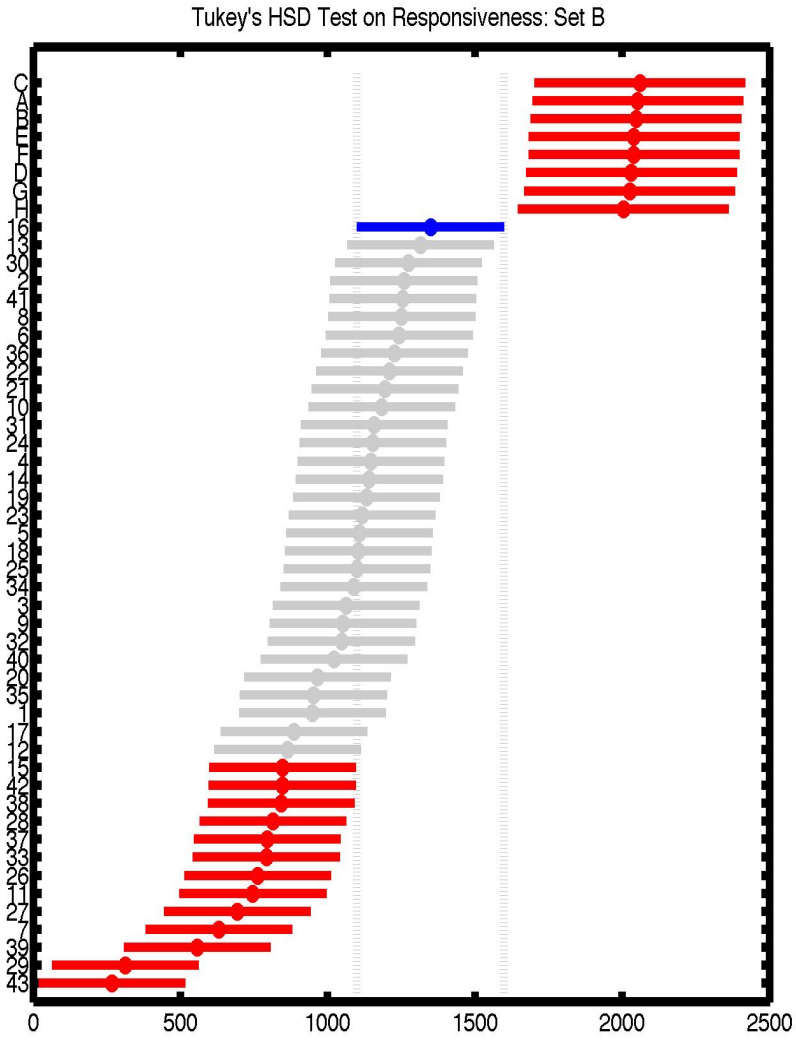
- Calculated from non-parametric ANOVA followed by Tukey honestly significant difference test

  NIST uses parametric ANOVA which assumes normality

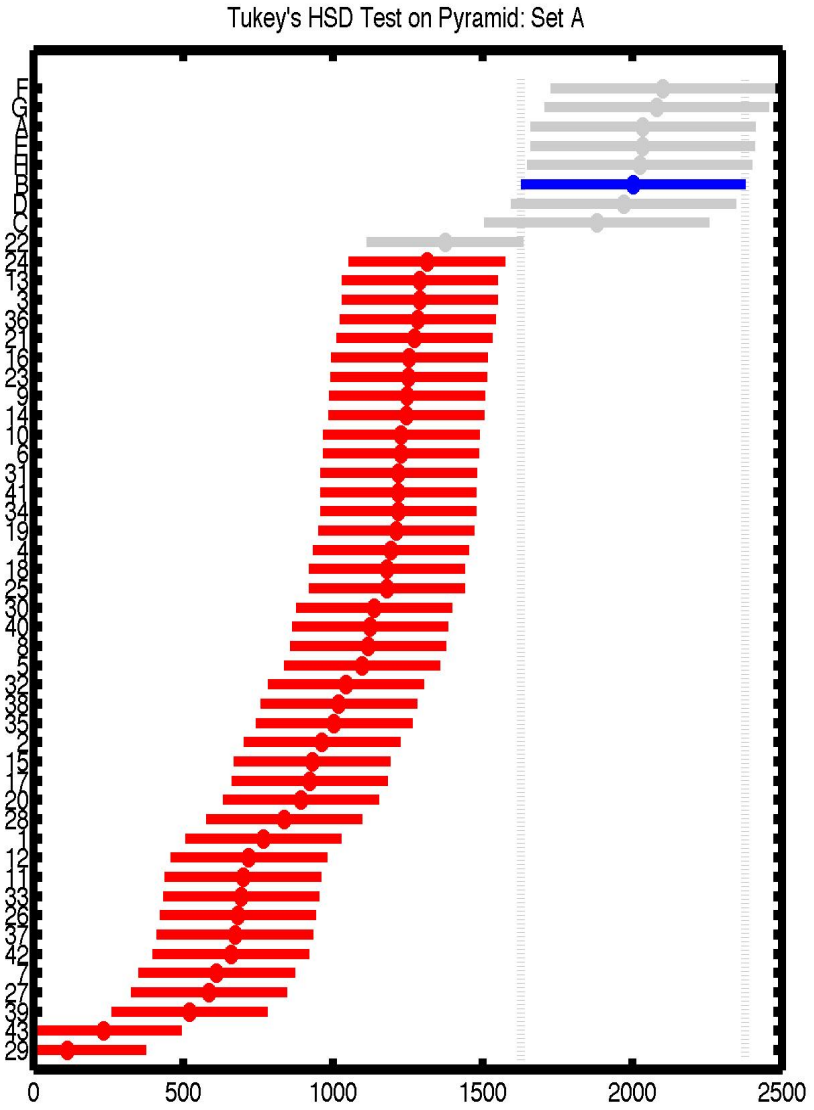
- We did well!

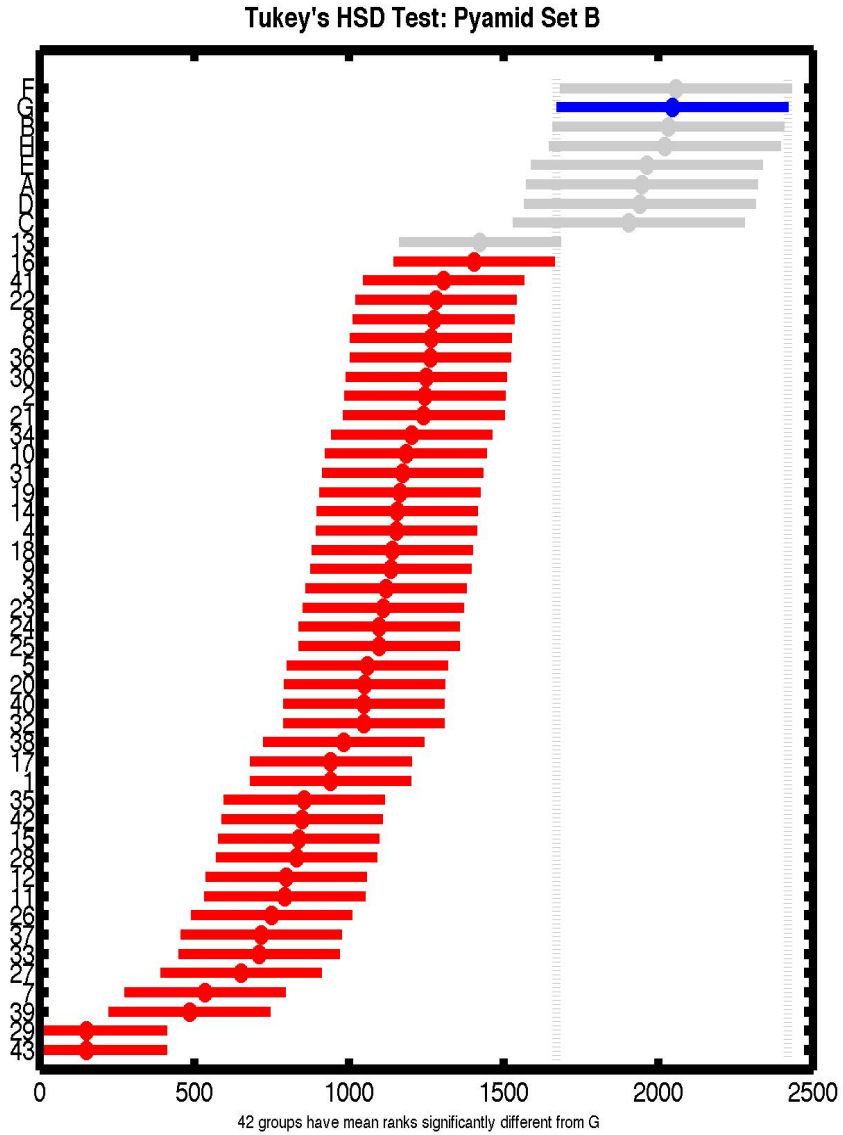
# Overall Responsiveness



# Pyramid Scores



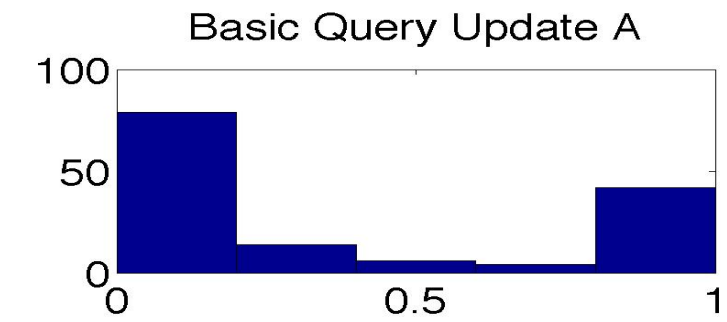



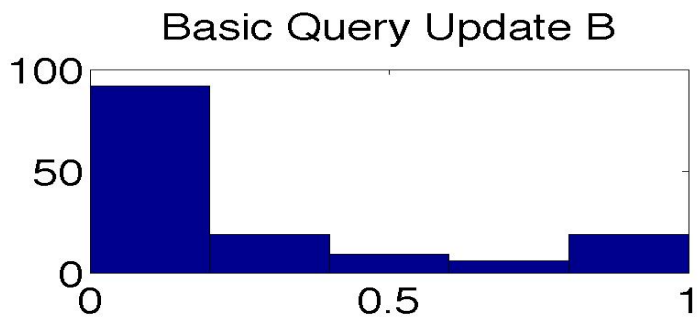
# Future Efforts

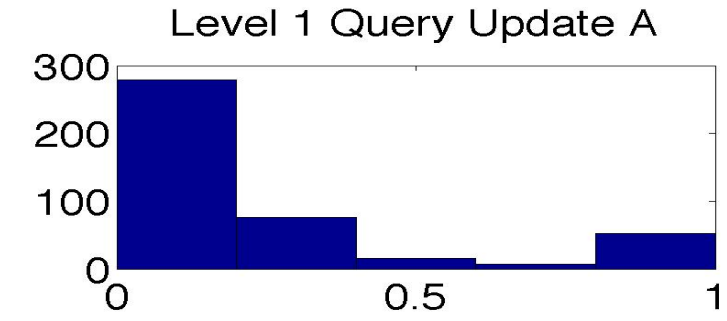
- Extend trimming capabilities
- Use more aspect information for sentence selection
- Continue to improve scoring, selection, and ordering

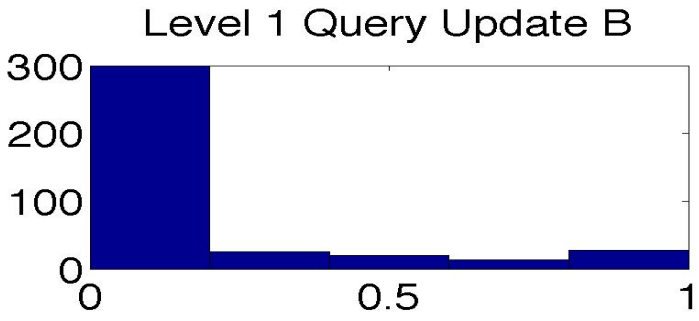
???

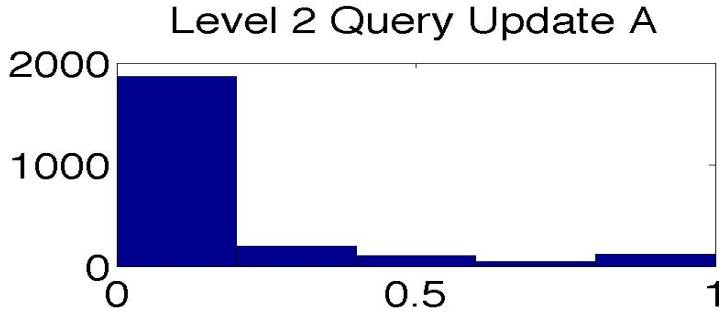
# Oracle Scores for Query Terms: Training Data

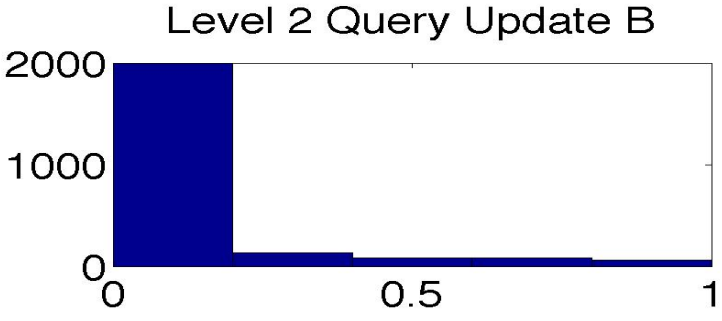


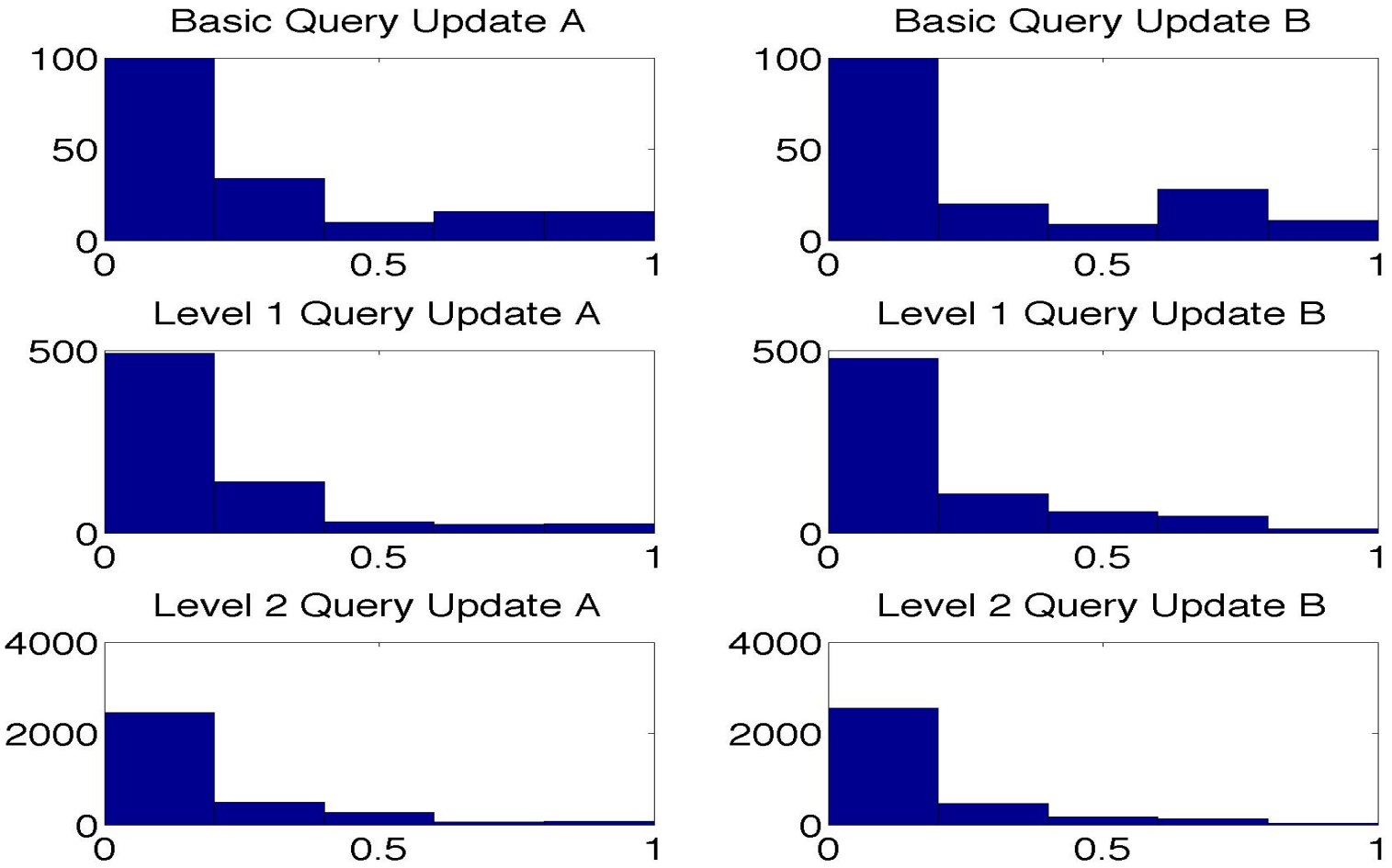



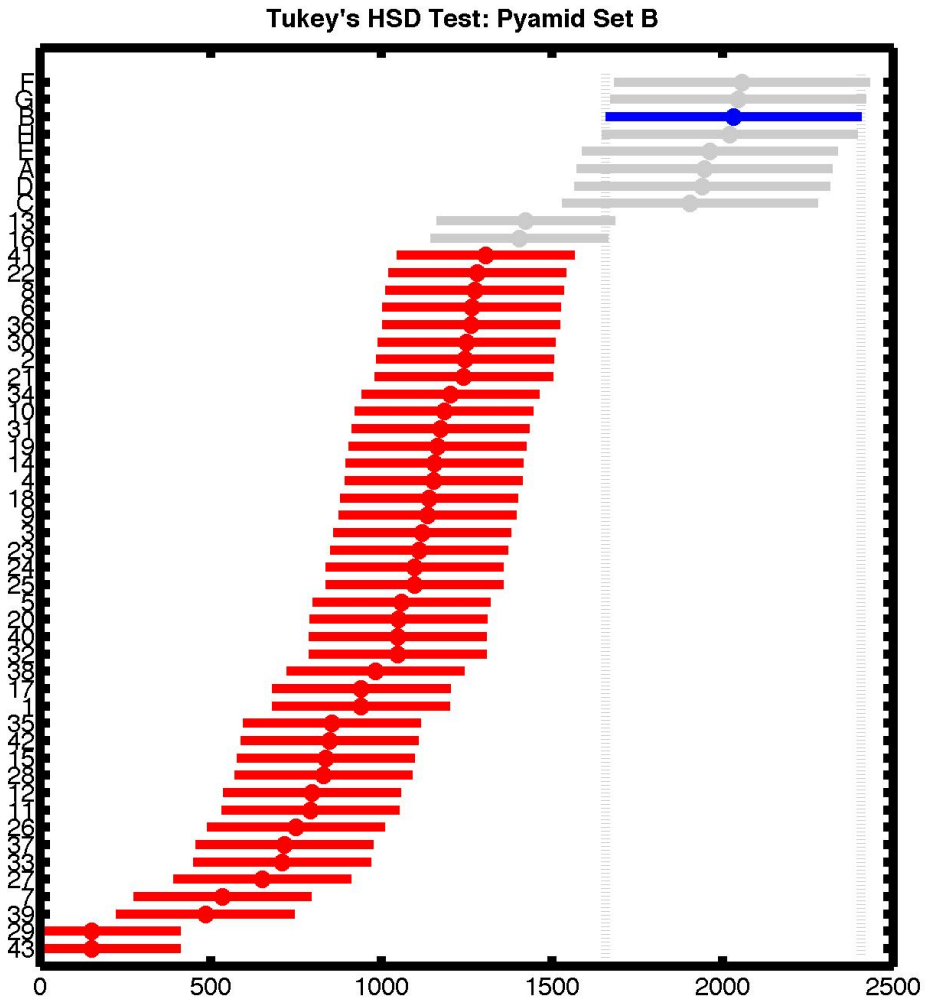



# Oracle Scores for Query Terms: TAC2010 Data




Tukey's HSD Test: Pyamid Set B
0
500
1000
1500
2000
2500

# Elimination Examples

- Relative Clause Appositive

  "Gen. Augusto Pinochet, who ruled Chile as a despot for 17 years, has been arrested in London after Spain asked that he be extradited for the presumed …."

- Attribution

  "… for around dlrs 137,000 to families of those killed in last month's crash off Nova Scotia, a spokesman for the airline said Tuesday."

# Gerund Phrase

Example:

"Suicide bombers targeted a crowded open-air market Friday, setting off several blasts."