# Overview of the 2008
# Text Analysis Conference

TAC

Sponsored by: NIST IARPA

BE THE FUTURE

Hoa Trang Dang
National Institute of Standards and Technology

# TAC 2008 Advisory Committee

Hoa Trang Dang *(NIST)*

John Burger *(MITRE)*

John Conroy *(IDA/CCS)*

Ido Dagan *(Bar Ilan U)*

Maarten de Rijke *(U Amsterdam)*

Bonnie Dorr *(U Maryland)*

Donna Harman *(NIST)*

Andy Hickl *(LCC)*

Ed Hovy *(ISI/USC)*

Boris Katz *(MIT)*

Bernardo Magnini *(ITC-irst)*

Kathy McKeown *(Columbia U)*

Ani Nenkova *(U Pennsylvania)*

Drago Radev *(U Michigan)*

Lucy Vanderwende *(Microsoft)*

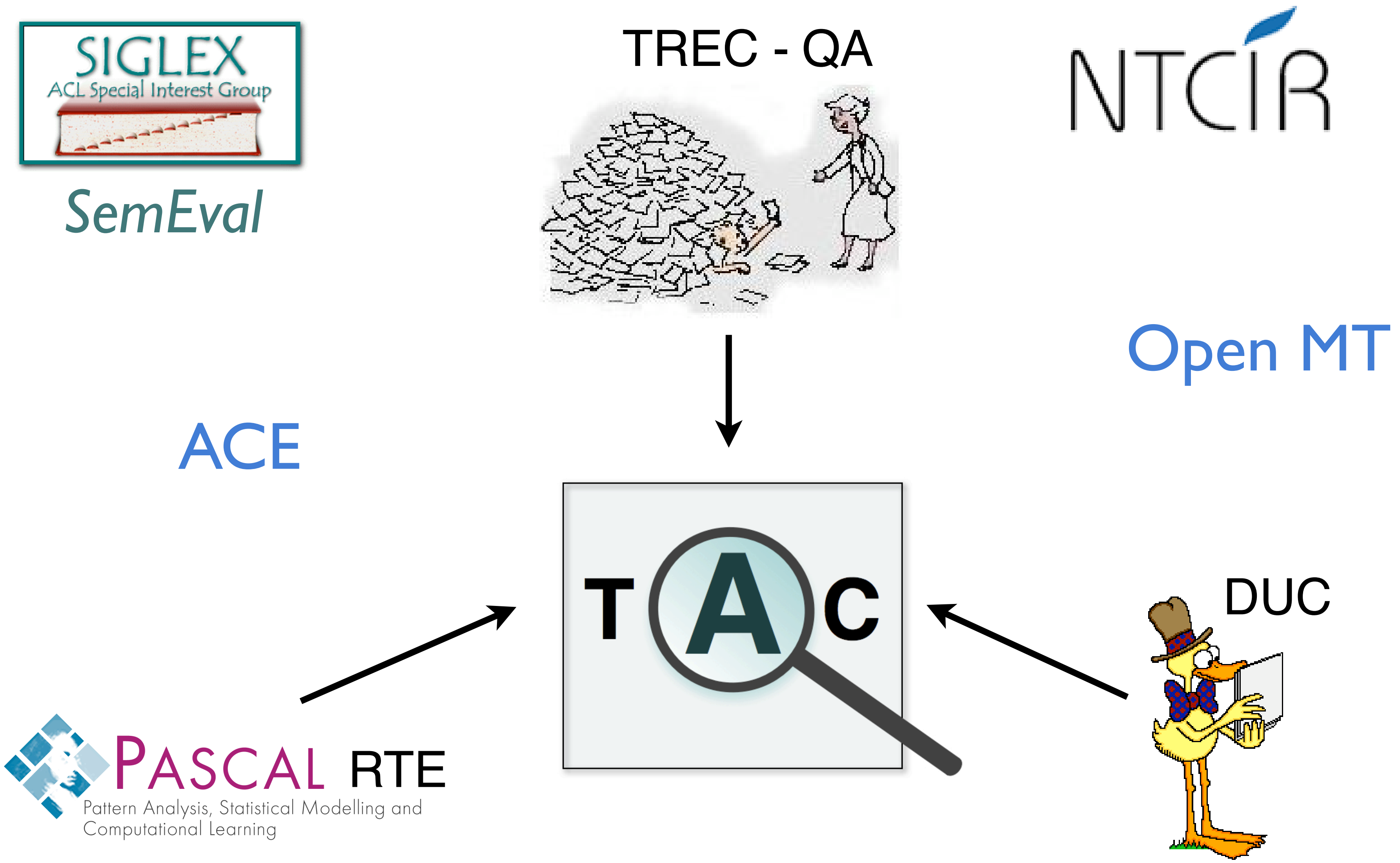
Ellen Voorhees *(NIST)*

Ralph Weischedel *(BBN)*

# Track Participants

- Track Organizers
  - Question Answering: Hoa Dang
  - RTE: Danilo Giampiccolo, Hoa Dang, Ido Dagan, Bernardo Magnini, Bill Dolan; *with support from Pascal-2 Network of Excellence*
  - Summarization: Hoa Dang *and Karolina Owczarzak*
- 65 Teams
  - 20 countries
  - 6 continents (23 N. America, 24 Europe, 15 Asia, ...)
- CELCT annotators, 17 NIST Assessors

# Why TAC?



SIGLEX

ACL Special Interest Group

SemEval

TREC - QA

NTCIR

Open MT

ACE

TAC

DUC

PASCAL RTE

Pattern Analysis, Statistical Modelling and Computational Learning

# TAC Goals

- To promote research in NLP based on large common test collections
- To improve evaluation methodologies and measures for NLP
- To build test collections that evolve to anticipate the evaluation needs of modern NLP systems
- To increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas
- To speed transfer of technology from research labs into commercial products

# Features of TAC

- Component evaluations situated within context of end-user tasks (e.g., summarization, QA)
  - opportunity to test components in end-user tasks
- Test common techniques across tracks
- Small number of tracks
  - critical mass of participants per track
  - sufficient resources per track (data, assessing, technical support)
- Leverage shared resources across tracks (organizational infrastructure, data, assessing, tools)

# TAC 2008 Tracks

- RTE: systems recognize when one piece of text entails or contradicts another
- QA: systems return a precise answer in response to a question, focusing on opinion questions asked over blogs
- Summarization: systems return a fluent summary of documents focused by a narrative or set of questions
  - Update: summarize new information in newswire articles for a user who has already read an earlier set of articles
  - Opinion pilot: summarize blog documents containing answers to opinion question(s) -- joint with QA

# Recognizing Textual Entailment (RTE)

- Goal: recognize when one piece of text is entailed by another
- Classification Task: given T(ext) and H(ypothesis)
  - H is entailed by T
  - H is not entailed by T
    - H contradicts T
    - H neither contradicts nor is entailed by T
- Optionally rank T-H pairs by entailment confidence
- Data collection from multiple application settings (IR, IE, QA, Summarization), correspond to success/failure cases in the settings

# RTE Pairs from IR Setting

- H: propositional IR queries
  - Becker was a tennis champion.
- T: extracted from documents retrieved by search engines (Google, Yahoo!, MSN, etc.)
  - Boris Franz Becker, German tennis player who, on July 7, 1985, became the youngest champion in the history of the men's singles at Wimbledon.

# RTE Pairs from IE Setting

- H: constructed from relations tested in ACE (plus new relations)
  - Ex-British Culture Minister suffers from AIDS.
- T: from text output by IE systems and humans performing IE task over news articles
  - Ex-British Culture Minister Chris Smith has lived with the HIV virus for 17 years, he has revealed to the Sunday Times. The former cabinet minister decided to go public after the former South African President Nelson Mandela announced his son Makgatho Mandela had died of AIDS.

# RTE Pairs from QA Setting

- H: generated from questions and candidate answer terms returned by QA systems searching the Web
  - *Baldwin* is Antigua's Prime Minister.
- T: candidate answer passages returned by QA systems
  - The opposition Antigua Labour Party (ALP) has blasted that country's prime minister, Baldwin Spencer, for publicly advocating that Cuba's Fidel Castro be awarded the Order of the Community (OCC) - the Community's highest honour.

# RTE Pairs from Summarization Setting

- H: simplified sentence from a set of news documents, or from multi-doc summary
  - Lady Bird Johnson, the first lady who championed conservation, died at 94.
- T: sentence(s) from a news document or (usually) summary
  - Lady Bird Johnson, the former first lady who championed conservation and worked tenaciously for the political career of her husband, Lyndon B. Johnson, died Wednesday, a family spokeswoman said. She was 94.

# RTE Evaluation

- Primary measure is accuracy:
  - 3-way accuracy (Entailed, Contradicts, Unknown)
  - 2-way accuracy (Entailed, Not Entailed)
- Secondary measure (for ranked pairs):
  - Average Precision

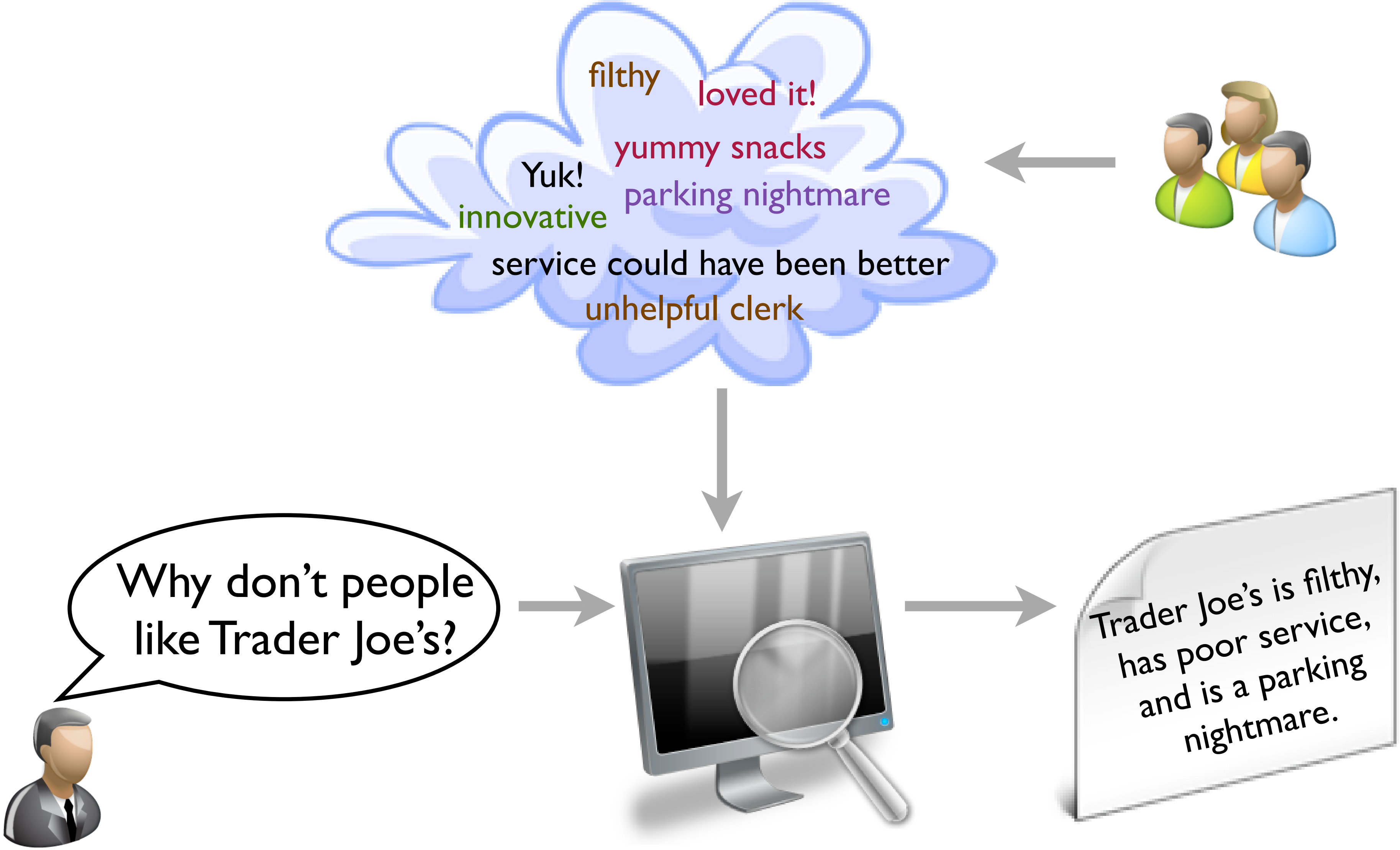
# Update Summarization Task

- Given a topic and 2 chronologically ordered clusters of news articles, A and B, where A documents precede B documents
- Create two brief (<=100 words), fluent summaries that contribute to satisfying the information need expressed in the topic statement:
  - Initial summary (A): summary of cluster A
  - Update summary (B): summary of cluster B, assuming reader has read cluster A

# Update Summary Evaluation

- Evaluation:
  - Pyramid Evaluation of summary content (Passonneau et al., 2005)
    - multiple human summaries
    - summary content unit (“nugget”) weighted by number of human summaries it appears in
  - Overall Readability
  - Overall Responsiveness (“What would I pay for this summary of the answers to my questions?”)
- For systems, update summaries more difficult than initial summaries (lower content and overall responsiveness scores)

# Pipelined Opinion QA/Summarization Task



filthy

loved it!

yummy snacks

Yuk!

parking nightmare

innovative

service could have been better

unhelpful clerk

Why don’t people like Trader Joe’s?

Trader Joe’s is filthy, has poor service, and is a parking nightmare.

# Opinion QA

TARGET: "MythBusters"

1018.1 RIGID LIST Who likes Mythbuster’s?

BLOG06-3334 CAPS_CHAMP
BLOG06-8580 Jon
BLOG06-3982 Zonk

1018.2 SQUISHY LIST Why do people like Mythbuster’s?

BLOG06-6706 The Mythbusters chicas are purdy .
BLOG06-5962 It's geek, period. And a lot of fun. I like that they have women on their team who are also into mechanical stuff and applied science.

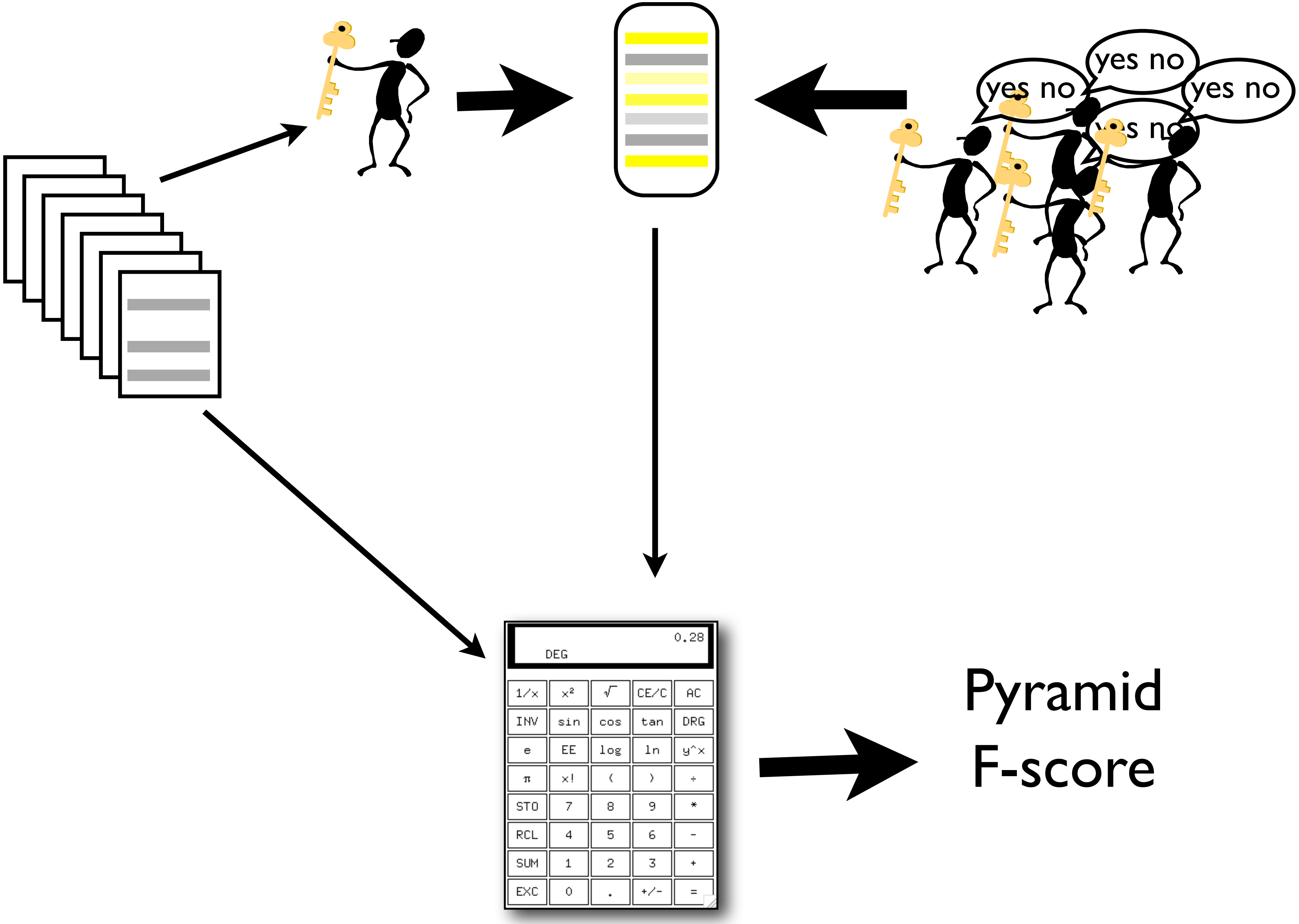
1018.3 RIGID LIST Who do people like on Mythbuster’s?

BLOG06-3187 Kari Byron
BLOG06-4849 scottie
BLOG06-6570 Jamie Hyneman

# Rigid List Evaluation

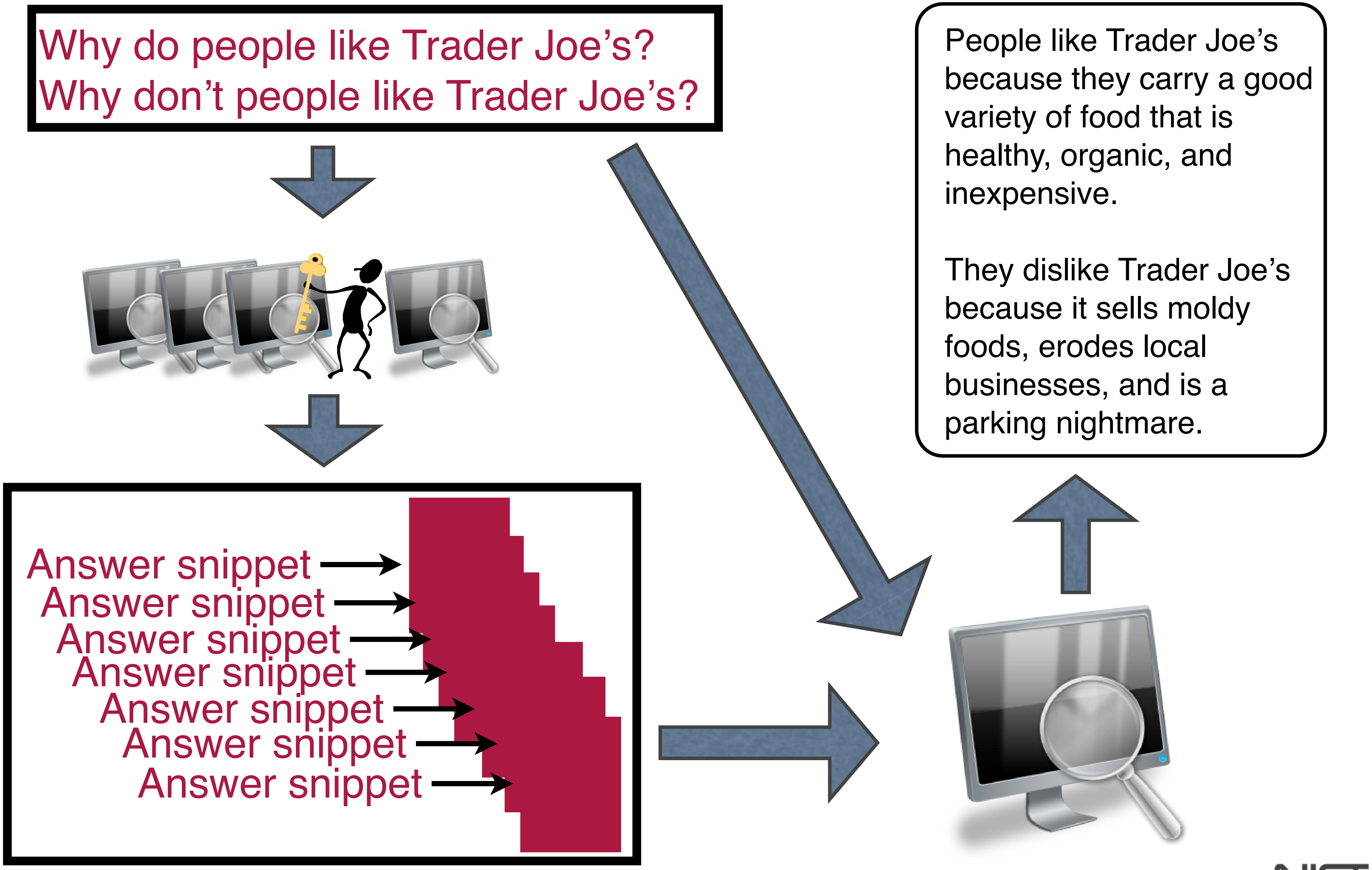
- Unordered list of [docid, answer-string] pairs
- Each pair judged as one of {wrong, unsupported, inexact, correct}
- Correct pairs grouped into equivalence classes (entities)
- Recall: number of correct entities returned / number of known correct entities
- Precision: number of correct entities returned / number of [docid, answer-string] pairs returned
- Combine precision and recall: F = (2*P*R)/(P+R)

# Squishy List Evaluation



Pyramid
F-score

# From Answer Snippets to Answer Summary



Why do people like Trader Joe’s?
Why don’t people like Trader Joe’s?

Answer snippet
Answer snippet
Answer snippet
Answer snippet
Answer snippet
Answer snippet
Answer snippet

People like Trader Joe’s because they carry a good variety of food that is healthy, organic, and inexpensive.

They dislike Trader Joe’s because it sells moldy foods, erodes local businesses, and is a parking nightmare.

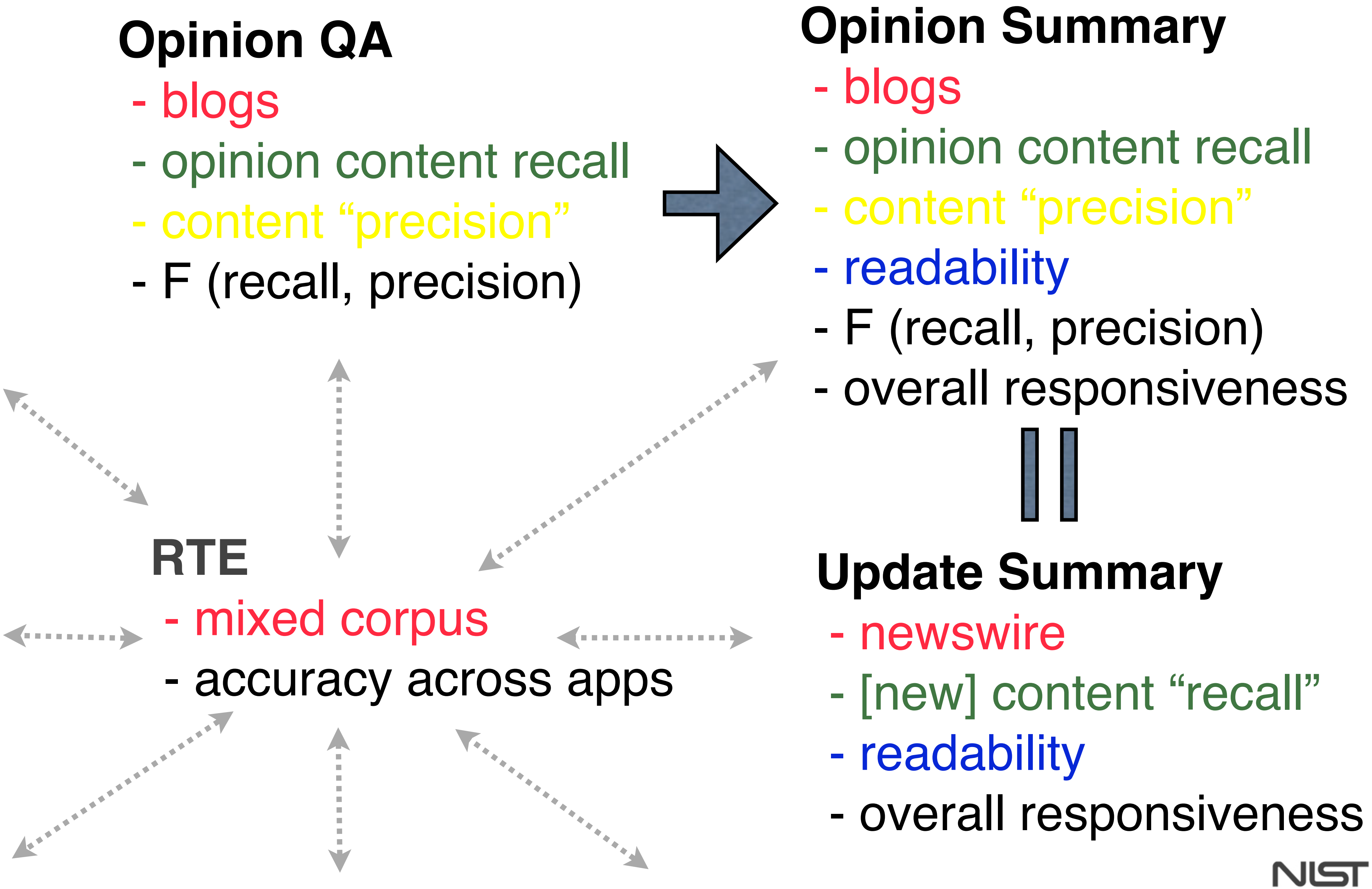
# Opinion Summarization

- Input
  - Target, 1-2 squishy list questions
  - Documents known to have answers
  - *Optional* answer-snippets in each document
- Output
  - Single fluent summary of the answers to all the questions
- Documents, snippets not labeled by question
- Opinion polarity classification (positive vs negative) may help fluency

# Opinion Summary Evaluation

- Pyramid F-score as for Opinion QA squishy list
- Readability score (grammaticality, non-redundancy, structure/coherence, overall readability)
- Overall Responsiveness score
- High correlation between average Pyramid F-score and overall responsiveness
- Low average readability scores (extraction from noisy blog text will be noisy)

# Common Threads: Applications + Components



**Opinion QA**

- blogs
- opinion content recall
- content “precision”
- F (recall, precision)

**Opinion Summary**

- blogs
- opinion content recall
- content “precision”
- readability
- F (recall, precision)
- overall responsiveness

=

**Update Summary**

- newswire
- [new] content “recall”
- readability
- overall responsiveness

**RTE**

- mixed corpus
- accuracy across apps

# Proposed TAC 2009 Tracks

1. RTE
2. Summarization:
   - Update ?
   - Opinion ?
   - Meeting/Speech ?
3. Information Extraction for Knowledge Base Population (successor to ACE)

- Come to the planning sessions!!