

# Summarization

**Hal Daumé III**

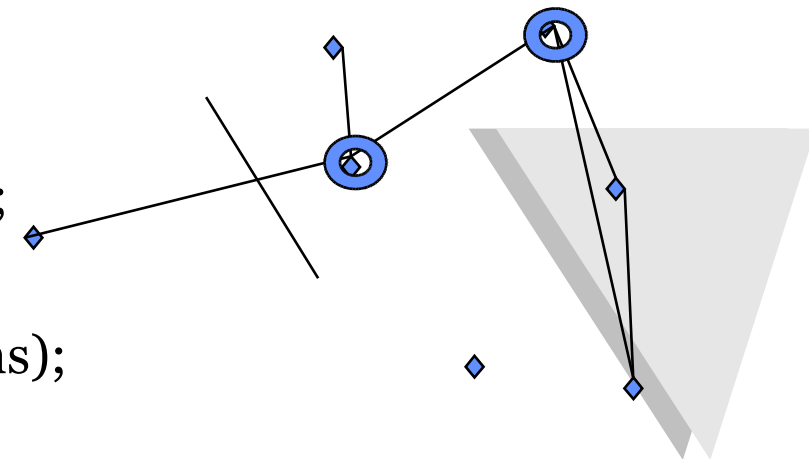
School of Computing  
University of Utah

me@hal3.name



# Cohesion-based methods

- **Claim:** Important sentences/paragraphs are the highest connected entities in more or less elaborate semantic structures.
- Classes of approaches
  - word co-occurrences;
  - local salience and grammatical relations;
  - co-reference;
  - lexical similarity (WordNet, lexical chains);
  - combinations of the above.



# Cohesion: Lexical chains method (1)

Based on (Morris and Hirst, 91)

But Mr. Kenny's move speeded up work on a **machine** which uses **micro-computers** to control the rate at which an *anaesthetic* is pumped into the blood of *patients* undergoing *surgery*. Such **machines** are nothing new. But Mr. Kenny's **device** uses two **personal-computers** to achieve much closer monitoring of the **pump** feeding the *anaesthetic* into the *patient*. Extensive testing of the **equipment** has sufficiently impressed the authorities which regulate *medical equipment* in Britain, and, so far, four other countries, to make this the first such **machine** to be licensed for commercial sale to *hospitals*.

# Discourse-based method

- **Claim:** The multi-sentence coherence structure of a text can be constructed, and the ‘centrality’ of the textual units in this structure reflects their importance.
- Tree-like representation of texts in the style of *Rhetorical Structure Theory* (Mann and Thompson, 88).
- Use the discourse representation in order to determine the most important textual units.  
Attempts:
  - (Ono et al., 94) for Japanese.
  - (Marcu, 97) for English.

# Rhetorical parsing

(Marcu,97)

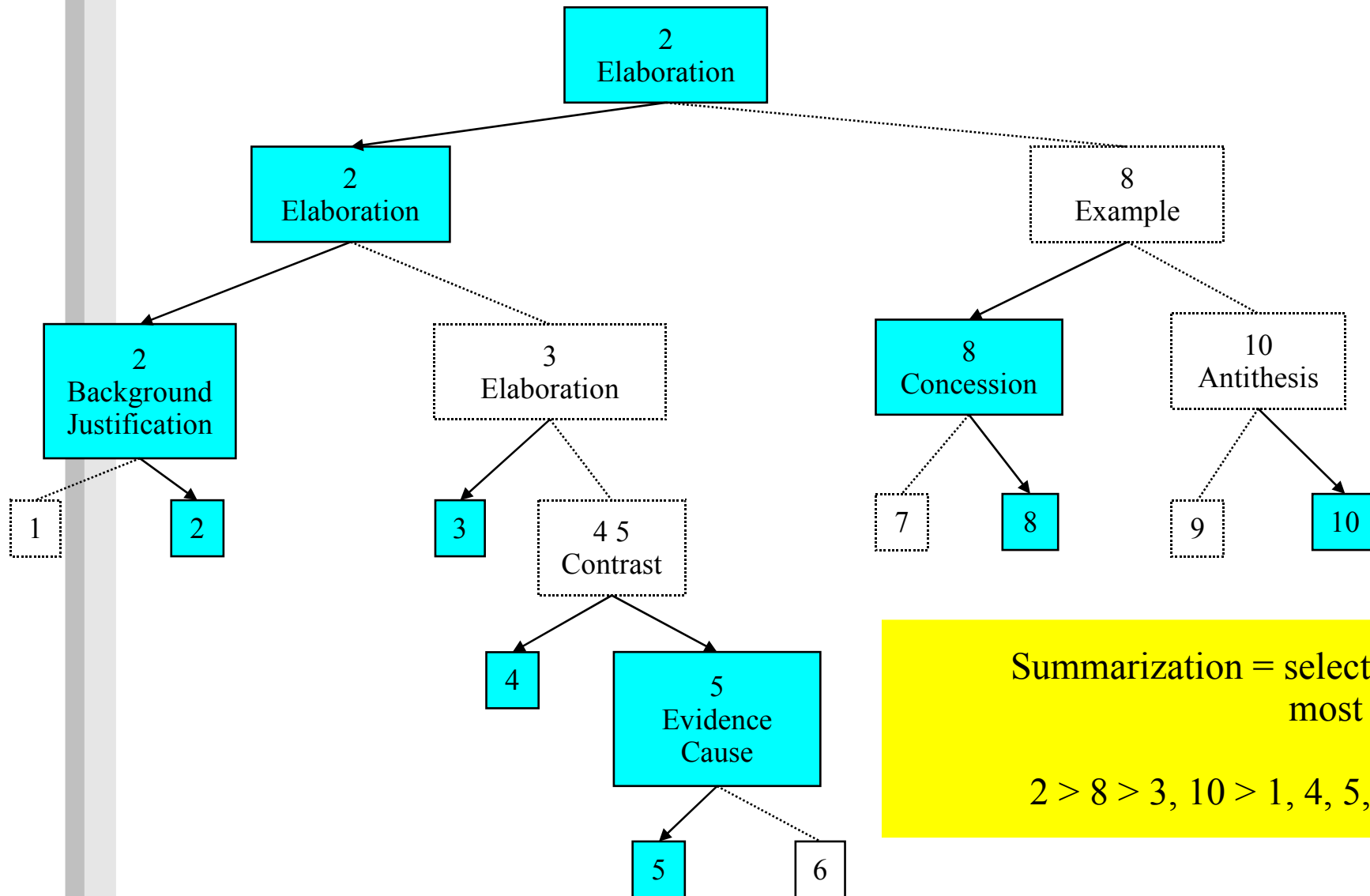
[*With* its distant orbit {– 50 percent farther from the sun than Earth –} and slim atmospheric blanket,<sup>1</sup>] [Mars experiences frigid weather conditions.<sup>2</sup>] [Surface temperatures typically average about –60 degrees Celsius (–76 degrees Fahrenheit) at the equator and can dip to –123 degrees C near the poles.<sup>3</sup>] [Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion,<sup>4</sup>] [*but* any liquid water formed that way would evaporate almost instantly<sup>5</sup>] [*because* of the low atmospheric pressure.<sup>6</sup>]

[*Although* the atmosphere holds a small amount of water, and water-ice clouds sometimes develop,<sup>7</sup>] [most Martian weather involves blowing dust or carbon dioxide.<sup>8</sup>] [Each winter, *for example*, a blizzard of frozen carbon dioxide rages over one pole, and a few meters of this dry-ice snow accumulate as previously frozen carbon dioxide evaporates from the opposite polar cap.<sup>9</sup>] [*Yet* even on the summer pole, {*where* the sun remains in the sky all day long,} temperatures never warm enough to melt frozen water.<sup>10</sup>]

## Rhetorical parsing (2)

- Use discourse markers to hypothesize rhetorical relations
  - $\text{rhet\_rel}(\text{CONTRAST}, 4, 5) \oplus \text{rhet\_rel}(\text{CONTRAT}, 4, 6)$
  - $\text{rhet\_rel}(\text{EXAMPLE}, 9, [7,8]) \oplus \text{rhet\_rel}(\text{EXAMPLE}, 10, [7,8])$
- Use semantic similarity to hypothesize rhetorical relations
  - if  $\text{similar}(u_1, u_2)$  then
    - $\text{rhet\_rel}(\text{ELABORATION}, u_2, u_1) \oplus \text{rhet\_rel}(\text{BACKGROUND}, u_1, u_2)$
    - else
    - $\text{rhet\_rel}(\text{JOIN}, u_1, u_2)$
  - $\text{rhet\_rel}(\text{JOIN}, 3, [1,2]) \oplus \text{rhet\_rel}(\text{ELABORATION}, [4,6], [1,2])$
- Use the hypotheses in order to derive a valid discourse representation of the original text.

# Rhetorical parsing (3)

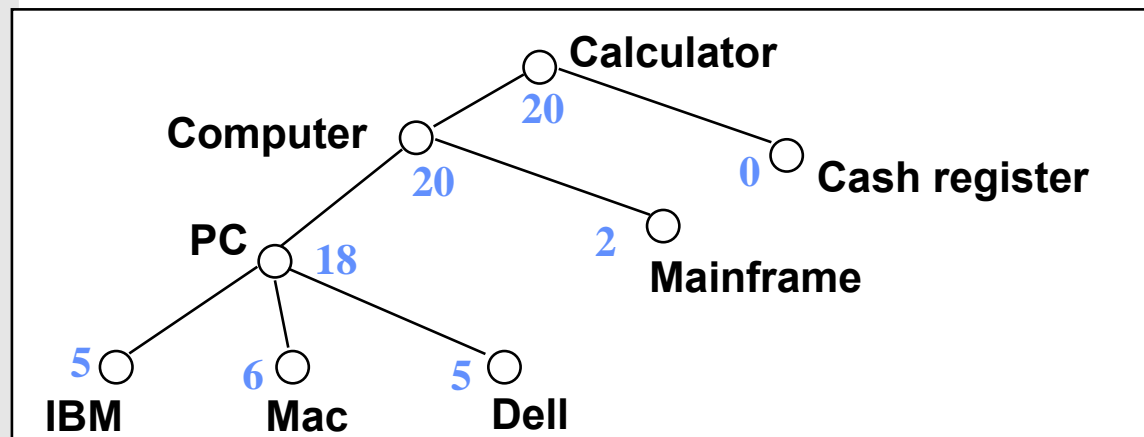


Summarization = selection of the most important units

$2 > 8 > 3, 10 > 1, 4, 5, 7, 9 > 6$

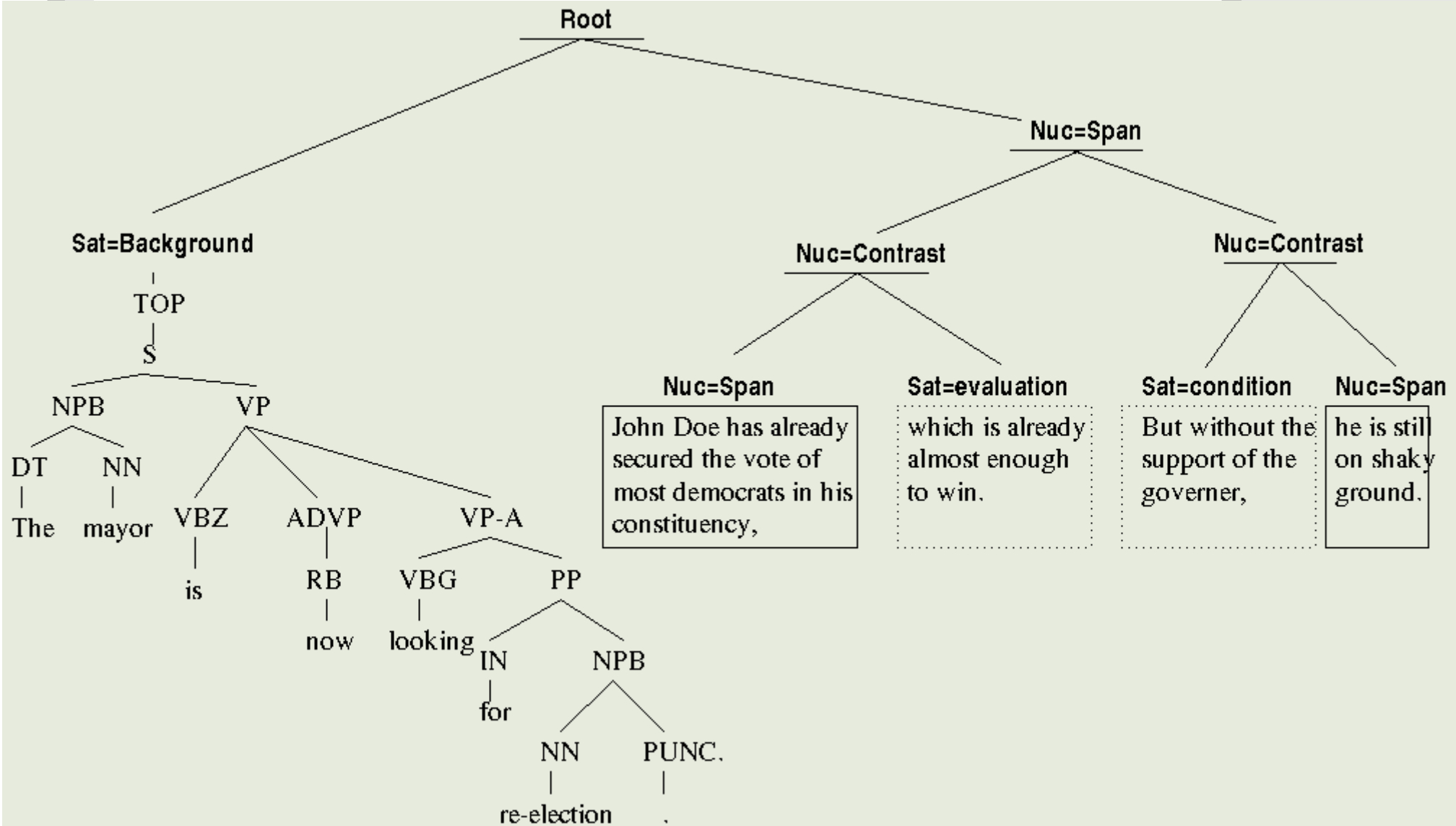
# Concept Generalization: Wavefront

- **Claim:** Can perform concept generalization, using WordNet (Lin, 95).
- Find most appropriate summarizing concept:



1. Count word occurrences in text; score WN concs
2. Propagate scores upward
3.  $R = \text{Max}\{\text{scores}\} / \Sigma \text{scores}$
4. Move downward until no obvious child:  $R < R_t$
5. Output that concept

# Compression model



[Knight+Marcu 2000, Daume+Marcu 2002]

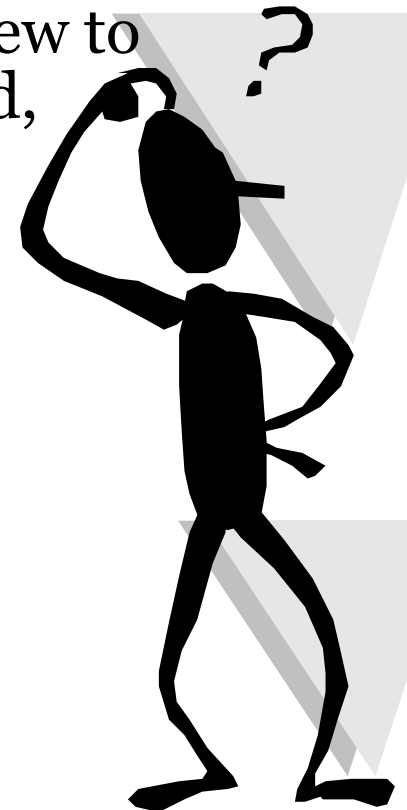
# What makes a good summary?

- (Only) Includes information relevant to user
- Is coherent
- Is cohesive
- Is grammatical (as appropriate for genre)
- Contains no dangling references
- Is terse

# How can You Evaluate a Summary?

- **When you already have a summary...**  
...then you can compare a new one to it:
  1. choose a granularity (clause; sentence; paragraph),
  2. create a similarity measure for that granularity (word overlap; multi-word overlap, perfect match),
  3. measure the similarity of each unit in the new to the most similar unit(s) in the gold standard,
  4. measure Recall and Precision.e.g., (Kupiec et al., 95).

..... **but when you don't?**



# Toward a Theory of Evaluation

## ➤ Two Measures:

Compression Ratio:  $CR = (\text{length } S) / (\text{length } T)$

Retention Ratio:  $RR = (\text{info in } S) / (\text{info in } T)$

## ➤ **Measuring length:**

- Number of letters? words?

## ➤ **Measuring information:**

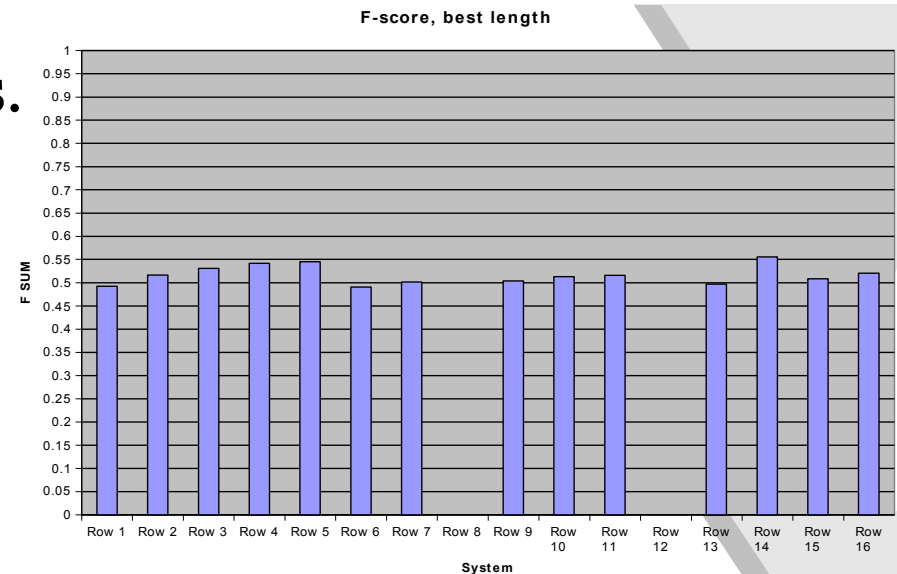
- *Shannon Game*: quantify information content.
- *Question Game*: test reader's understanding.
- *Classification Game*: compare classifiability.

# SUMMAC Categorization Test

- **Procedure** (SUMMAC, 98):
  1. 1000 newspaper articles from each of 5 categories.
  2. Systems summarize each text (generic summary).
  3. Humans categorize summaries into 5 categories.
  4. Testers measure *Recall* and *Precision*, combined into *F*:  
*How correctly are the summaries classified, compared to the full texts?*  
(many other measures as well)

- **Results:**

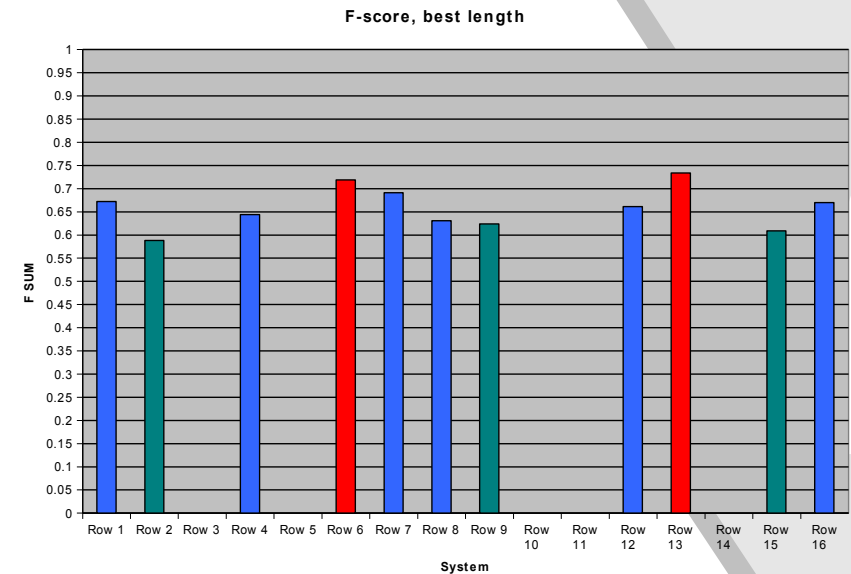
No significant difference!



# SUMMAC Ad Hoc (Query-Based) Test

- **Procedure** (SUMMAC, 98):
  1. 1000 newspaper articles from each of 5 categories.
  2. Systems summarize each text (query-based summary).
  3. Humans decide if summary is relevant or not to query.
  4. Testers measure  $R$  and  $P$ : *how relevant are the summaries to their queries?* (many other measures as well)

- **Results:**  
3 levels of performance



# ROUGE: Recall-based Evaluation

## Reference summary:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

Tri-gram match

Bi-gram matches

## System summary:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

“Rouge” metric