CLASSY Summarization-- English and Beyond

Judith D. Schlesinger
John M. Conroy
IDA Center for Computing Sciences
Joint Work with
Jeff Kubina, DOD
Dianne P. O'Leary, University of Maryland

Overview

- Linguistic Processing
  - Guided Summarization
  - Multi-lingual Summarization
  - Future Tasks
- Scoring and Selection
  - Guided Summarization
  - Multi-lingual Summarization
  - Future Tasks

Guided Summarization Linguistic Processing

- Tasks
  - Classify sentences: -1, 0, 1
  - Sentence split: FASST-E
  - Tokenize and trim
  - Query term generation

Guided Summarization Linguistic Processing (cont.)

- Basically very stable
  - Changing only to correct errors or to handle new situations
- But …
  - Error in “clean” data
  - Others

Multi-lingual Summarization Linguistic Processing

- New: 2 variations for other languages
  - Based on FASST-E
  - upper/lower case alphabets; single case only
  - Growing pain errors
    - Missed splits after numbers
- New formats...new problems
  - Datelines, including English
  - Catch-22 on how to handle

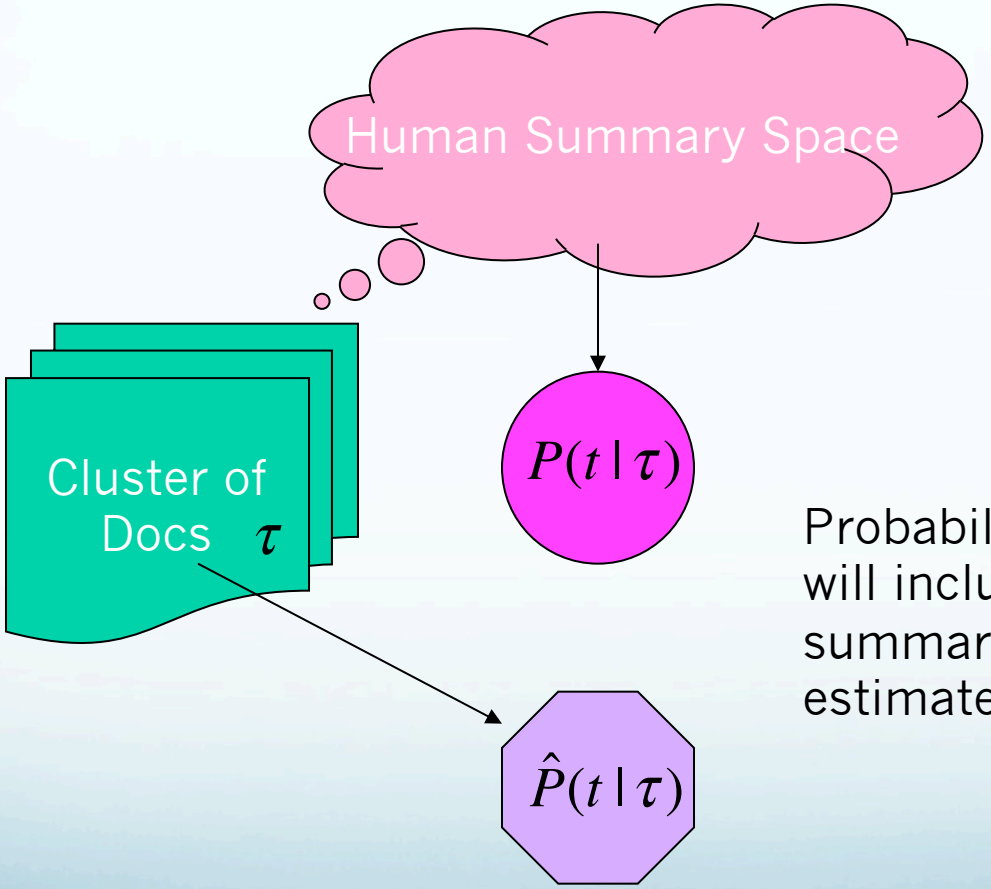
Linguistic Processing Future Tasks

- Strengthen non-English sentence splitters
  - 2nd pass for datelines, quotes, short sentences, etc.
- Non-English trimming
  - Lead phrases
  - Other trims????
- English: Anaphora resolution

Questions???

- Examples of new dateline formats
  - Tuesday, July 18, 2005
  - Meadow Lake, Saskatchewan --
  - On same line as following text

Mathematical Model



Probability that a human will include term *t* in a summary on topic τ and an estimate.

General Recipe

1. Estimate probability that a term (bigram) will be included by a human.
2. Optionally project term sentence matrix to be orthogonal to previously generated summary.
3. Select a non-redundant subset of sentences with high density of terms likely chosen by a human.
4. Order the sentences to improve flow (approximate TSP).

Submission 25

$$P_{qs\rho}(t \mid \tau) = \alpha_q q(t) + \alpha_s s(t) + \alpha_\rho \rho(t)$$

$$s(t)[q(t)] = \begin{cases} 1 \text{ if } t \text{ is a signature [query] term} \\ 0 \text{ if } t \text{ is not a signature [query] term} \end{cases}$$

$$\rho(t \mid \tau) = \text{probability } t \text{ occurs in a sentence considered for selection.}$$

Followed by non-negative QR, knapsack to insure 100 words or less, and the approximate TSP to improve flow.
Major changes: bigrams and expanded query set.
Parameters set optimizing using ROUGE-2 and ROUGE-SU4 as well as nouveu variants for updates.

Submission 42

$$P_{\mathrm{NB}}(t \mid \tau) = \sum_{i=0}^{4} \tfrac{i}{4} P(i \mid f_1, f_2)$$

$P(i \mid f_1, f_2) =$ Bayes posterior prob that i humans would include a term whose features are f_1 and f_2.

Intitial Summaries:

$f_{11}^{A} = \log(p - \text{value used in signature term computation}$

$f_2^{A} =$ TextRank of term t.

Update Summaries: $f_1^{B} = \log(f_2^{B} / f_2^{A})$.

Low scoring non-query terms removed to compute TextRank.

Followed by non-negative QR, knapsack to insure 100 words or less, and

an approximate TSP to improve flow.

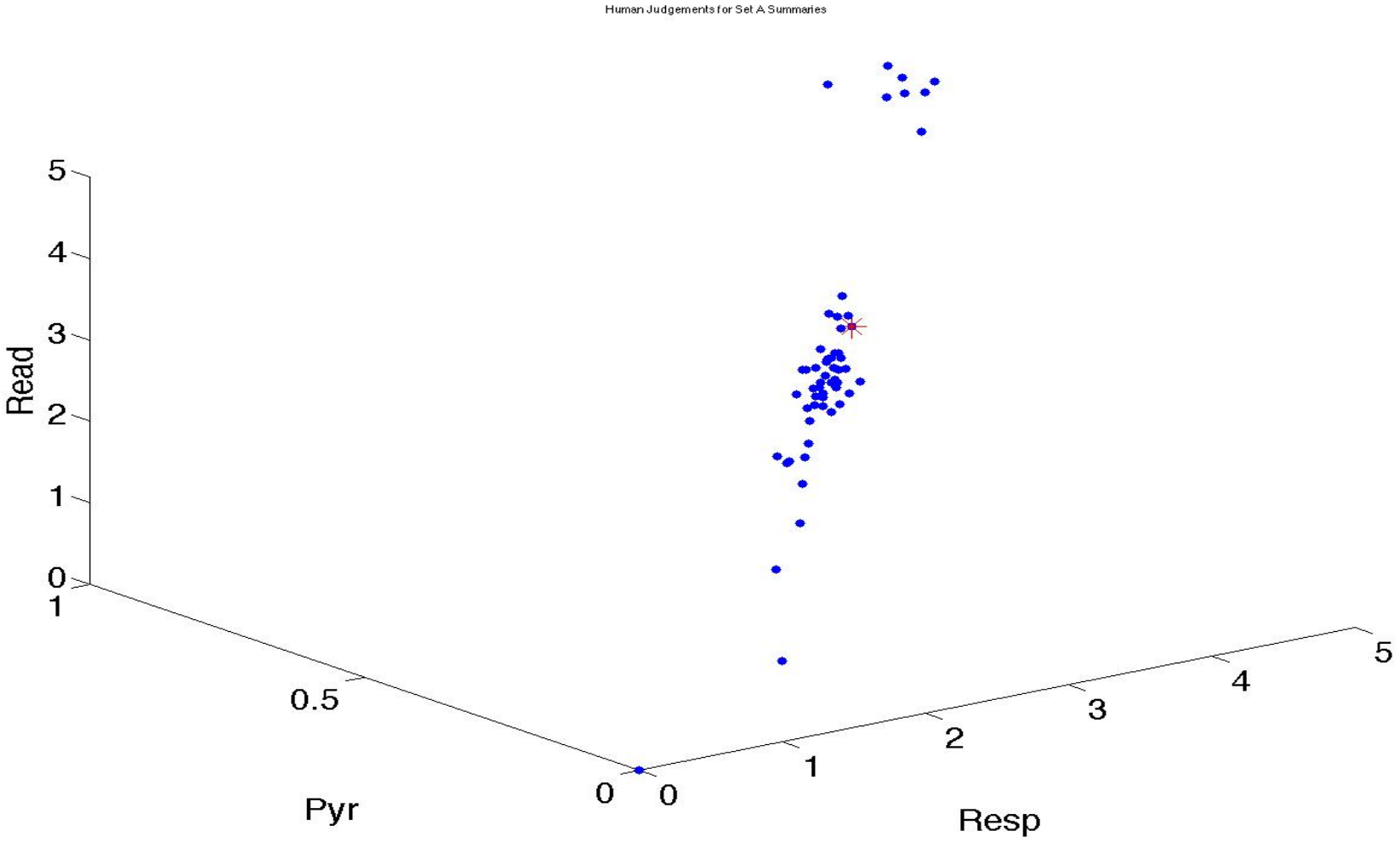
Major changes: bigrams and expanded query set.

Trained on TAC 2010 using naïve Bayes, normal approximation.

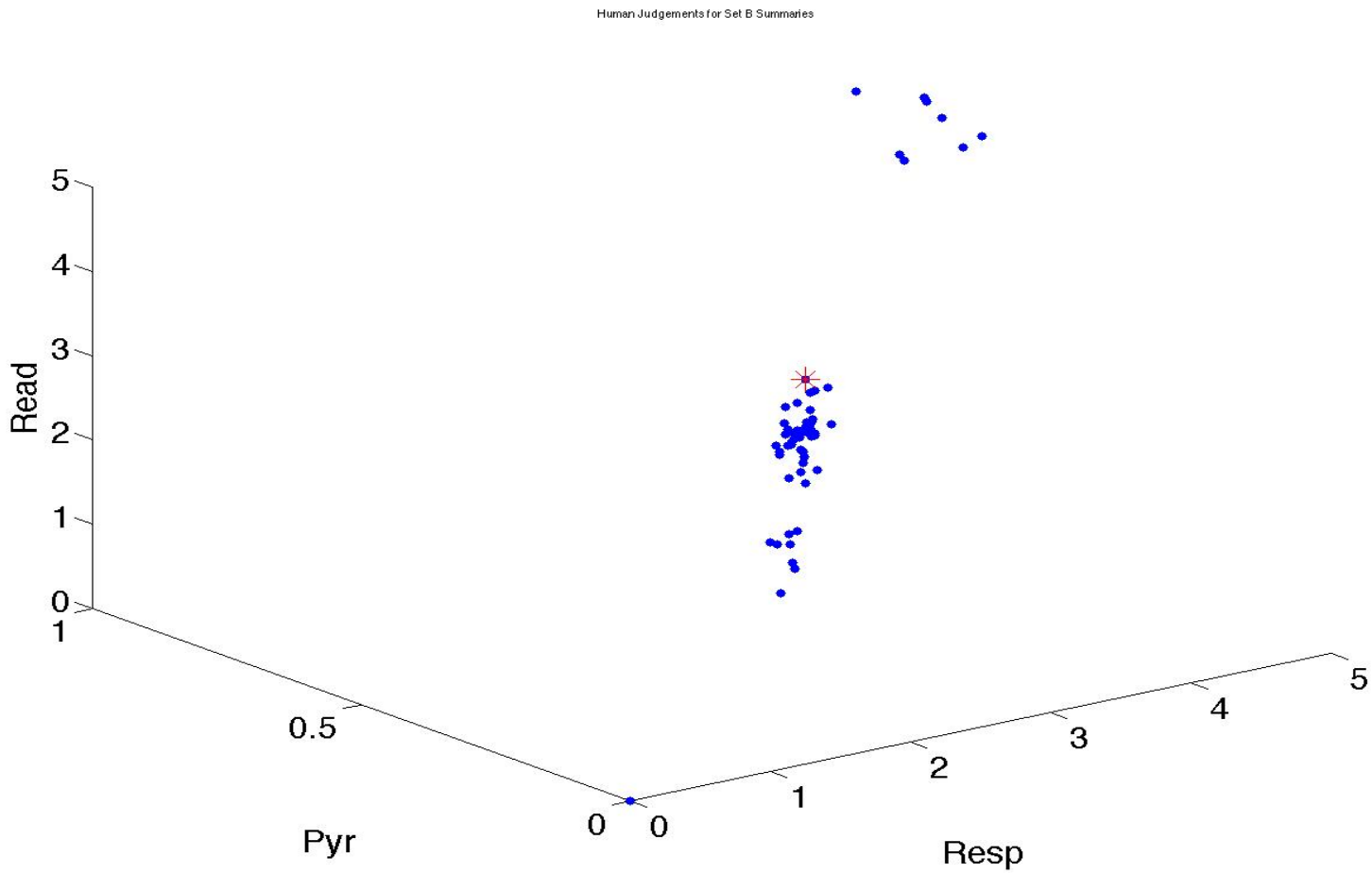
Results

Submission	Resp.	Pyr.	Read.	ROUGE-2 Rank (#humans beat)
25 Set A	1	10	6	3 (7)
25 Set B	3	4	2	2 (4)
42 Set A	18	28	9	9 (5)
42 Set B	17	26	9	15 (1)

A View of the Results



View of the Update Results



Multi-lingual Task

Goal: Develop a language independent summarizer.

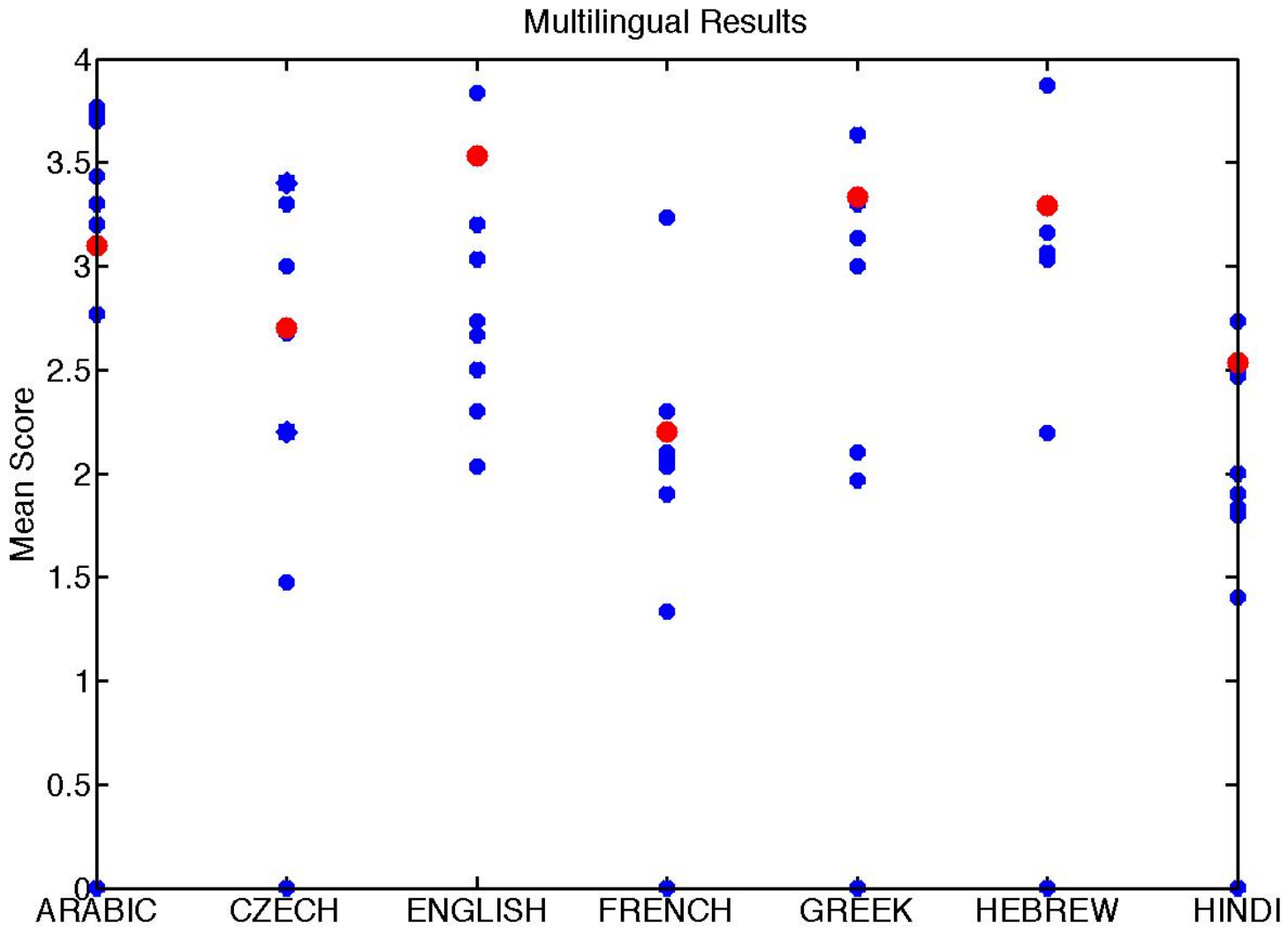
Approach:

1. Collect a background model for each target language(Wiki news).
2. Compute language independent features.
3. Train a naïve Bayes classifier on DUC 2005-2007 to compute $P_{NB}(t|\ \tau\ )$
4. Use binary integer linear program to achieve a maximum covering (better than non-negative QR > 100 words).

Features

1. log(p) p-value of Dunning (signature term) G-statistic.
2. Sentence TextRank; terms with p-value<0.001 are included. (Auto-stop list.)
3. log($P(t_j|S_0)$); log probability that a term occurs in a sentence in the cluster of documents to be summarized.
4. log($P(t_j|S_1)$); log probability that a term occurs in a sentence with 1 or more signature term in the cluster of documents to be summarized.

Multilingual Results



Things to Do

- Investigate further why ML failed to do as well.
- Investigate to what extent current features are language independent.
- Further use of pairwise testing to determine best approach. (See Peter Rankel's talk.)