Better Metrics to Automatically Predict the Quality of a Text Summary

Peter Rankel, University of Maryland
John M. Conroy, IDA Center for Computing Sciences
Judith D. Schlesinger, IDA Center for Computing Sciences
Why Automatically Evaluate Text Summaries?

• Create a useful tool for developers of automatic summarizers
• Other uses (possibly in the distant future....):
  – Human-produced summaries
  – Abstracts of technical papers
Text Analysis Conference

• Sponsored by NIST 2008-2011
• Summarization Task
  – 100-word summaries of sets of 20 newswire articles for given topics (10 and 10)

Initial Summary                         Update Summary
44 document sets

• Categories:
  – Accidents and Natural Disasters
  – Attacks (Criminal/Terrorist)
  – Health and Safety
  – Endangered Resources
  – Investigations and Trials (Criminal/Legal/Other)

• Each has particular aspects
### Aspects for:
Accidents and Natural Disasters

- **WHAT**: what happened
- **WHEN**: date, time, other temporal markers
- **WHERE**: physical location
- **WHY**: reasons for accident/disaster
- **WHO_AFFECTED**: casualties (death, injury), or individuals otherwise negatively affected
- **DAMAGES**:
- **COUNTERMEASURES**: rescue efforts, prevention efforts, other reactions
How NIST evaluates summaries

• 3 Human-produced scores
  – Pyramid score – count number of information “nuggets” in summary and compare to those in human-written summaries
  – Readability – Grammaticality, non-redundancy, referential clarity, focus, structure/coherence
  – Overall responsiveness – measures linguistic quality and how informative the summary is

• We would like to do this automatically
A.E.S.O.P.: Automatically Evaluating Summaries Of your Peers

• Create an algorithm to predict human measures of a summary’s quality

• Input:
  – Machine-produced summary
  – Four human-written summaries (“models”)

• Output:
  – A score that hopefully correlates well with any of the three human measures

• Compare our results to ROUGE-2, ROUGE-SU4
TAC 2011 Automatic vs overall with Linear Prediction.

- $\rho_{R2} = 0.92$
- $\rho_{SU4} = 0.94$
Our Algorithm(s)

• Combine subsets of 6 content features and 7 linguistic features using various forms of linear regression

• Regression techniques (Matlab)
  – Robust regression
  – Non-negative least squares
  – Canonical correlation
Features / Predictors

• Content Features
  – ROUGE-2
  – ROUGE-SU4
  – Bigram coverage score
  – Point-to-point versions of the above

• Linguistic Features
  – Number of sentences
  – 2 redundancy scores based on sums of squares of singular values
  – Term entropy
  – Sentence entropy
  – 2 term-overlap scores
Training the Models

• We have lots of training data.....
• Consider all subsets of our 13 features
• Use TAC 2009 data to select features
• Use TAC 2010 data to calculate coefficients
• Predict scores of summaries for TAC 2011
### Table 2: Features Used When Predicting Scores of both Human and Machine Summaries

<table>
<thead>
<tr>
<th>Feature</th>
<th>Pyramid(8)</th>
<th>Responsiveness (25)</th>
<th>Readability (23)</th>
<th>Responsiveness(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>canon</td>
<td>canon</td>
<td>robust</td>
<td>robust</td>
</tr>
<tr>
<td>R2</td>
<td>4.8e+1</td>
<td>4.8e+1</td>
<td>4.5e+1</td>
<td>2.3e+1</td>
</tr>
<tr>
<td>SU4</td>
<td>3.7e+1</td>
<td>3.7e+1</td>
<td>4.0e+1</td>
<td>1.2e-1</td>
</tr>
<tr>
<td>Coverage</td>
<td>3.6e-1</td>
<td>3.6e-1</td>
<td>2.7e-1</td>
<td>7.4e-2</td>
</tr>
<tr>
<td>Bigrams</td>
<td>-4.0e-1</td>
<td>-4.0e-1</td>
<td>-3.6e-1</td>
<td>3.5e-3</td>
</tr>
<tr>
<td>Bigrams P2P</td>
<td>5.9e-1</td>
<td>5.9e-1</td>
<td>3.0e+0</td>
<td>7.0e+2</td>
</tr>
<tr>
<td>Coverage P2P</td>
<td>-7.2e-1</td>
<td>-3.2e+0</td>
<td>-3.0e-5</td>
<td>2.9e-2</td>
</tr>
<tr>
<td>log2(1+Term Overlap)</td>
<td>8.9e-2</td>
<td>8.9e-2</td>
<td>7.6e-2</td>
<td>0.0e+0</td>
</tr>
<tr>
<td>Norm Term Overlap</td>
<td>-6.6e-1</td>
<td>-6.6e-1</td>
<td>-2.0e-4</td>
<td>6.7e-4</td>
</tr>
<tr>
<td>Redundant 1</td>
<td>-1.0e+0</td>
<td>-5.3e-2</td>
<td>-5.3e-2</td>
<td>2.2e-1</td>
</tr>
<tr>
<td>Term Entropy</td>
<td>1.2e+0</td>
<td>1.2e+0</td>
<td>2.2e-3</td>
<td>1.1e-2</td>
</tr>
<tr>
<td>-log2(sent length)</td>
<td>-9.6e-1</td>
<td>-9.6e-1</td>
<td>1.6e-1</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3: Features Used When Only Predicting Scores of Machine Summaries

<table>
<thead>
<tr>
<th>Feature</th>
<th>Pyramid(8)</th>
<th>Responsiveness (25)</th>
<th>Readability (23)</th>
<th>Responsiveness(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>robust</td>
<td>robust</td>
<td>robust</td>
<td>robust</td>
</tr>
<tr>
<td>R2</td>
<td>1.2e+1</td>
<td>2.1e+1</td>
<td>2.0e+0</td>
<td>2.0e+0</td>
</tr>
<tr>
<td>SU4</td>
<td>-5.7e-1</td>
<td>3.4e+1</td>
<td>4.2e+1</td>
<td>9.6e-3</td>
</tr>
<tr>
<td>Coverage</td>
<td>3.0e-1</td>
<td>9.0e+1</td>
<td>1.0e+2</td>
<td>5.1e-3</td>
</tr>
<tr>
<td>Bigrams</td>
<td>2.2e+1</td>
<td>3.0e-1</td>
<td>4.0e-1</td>
<td></td>
</tr>
<tr>
<td>Bigrams P2P</td>
<td>-8.6e+0</td>
<td>8.7e+0</td>
<td>6.8e+0</td>
<td></td>
</tr>
<tr>
<td>Coverage P2P</td>
<td>9.3e-2</td>
<td>1.4e-1</td>
<td>1.0e-1</td>
<td>1.0e-1</td>
</tr>
<tr>
<td>log2(1+Term Overlap)</td>
<td>-1.2e+0</td>
<td>-7.4e-1</td>
<td>-1.5e-4</td>
<td>6.0e-4</td>
</tr>
<tr>
<td>Norm Term Overlap</td>
<td>5.4e-4</td>
<td>4.6e-5</td>
<td>3.2e-4</td>
<td></td>
</tr>
<tr>
<td>Redundant 1</td>
<td>-1.0e-1</td>
<td>-1.0e-1</td>
<td>8.7e-2</td>
<td></td>
</tr>
<tr>
<td>Term Entropy</td>
<td>1.8e+0</td>
<td>5.2e+1</td>
<td>6.7e-2</td>
<td></td>
</tr>
<tr>
<td>-log2(sent length)</td>
<td>-2.0e+0</td>
<td>-4.6e-1</td>
<td>8.4e-2</td>
<td></td>
</tr>
<tr>
<td>Neg Sent Entropy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Our systems are numbers 25, 8, 23, and 6, finishing 2\textsuperscript{nd}, 3\textsuperscript{rd}, 4\textsuperscript{th}, and 6\textsuperscript{th} out of 25 submissions.
References

Thank You

• Contact info:
  – Peter Rankel, University of Maryland, rankel@math.umd.edu
  – John Conroy, IDA Center for Computing Sciences, conroy@super.org
  – Judith Schlesinger, IDA Center for Computing Sciences, judith@super.org