DUC 2006 Pyramid Evaluation

Rebecca J. Passonneau
Center for Computational Learning Systems
Columbia University
Acknowledgments

- Hoa Dang
- Columbia University (Kathy McKeown)
- Guideline contributors, testers (Lucy Vanderwende, Adam Goodkind, Guy LaPalme, . . .)
- Pyramid Creators (Adam Goodkind, Sergey Sigelman, Lucy Vanderwende, Inderjeet Mani, Qui Long)
- Participants (21 sites)
Pyramid Overview

- Human summarizers select overlapping content
- A pyramid represents and quantifies the overlap of Summary Content Units (SCUs) found in multiple model summaries
- Two pyramid scores based on SCU annotations
  - Original ≈ Precision
  - Modified ≈ Recall
- Manual annotation reliability assessment
  - Pyramid annotations (LREC 2006)
  - Peer annotations (DUC 2005)
Sample SCU from D0631

[Label: The Concorde crossed the Atlantic in less than 4 hours]

**Sum1** < making the transatlantic flight in 3 and ½ hrs >

**Sum2** < The Concorde could make the flight in between New York and London or Paris in less than four hours>

**Sum3** < completing its journey from London to New York in about 3 hours, 30 minutes >

**Sum4** < took less than 4 hrs to cross the Atlantic >

June 8, 2006  DUC Workshop
Building a Pyramid from Model Summaries (N=4)

W=4
W=3
W=2
W=1

June 8, 2006
2006 Pyramid effort

- New version of DUCView, annotation guidelines
- Pyramids for 20 of the document sets
  - High clarity ratings
  - Even distribution of assessors (summary writers)
- Pyramid annotation
  - 6 individuals at 3 sites, 2 with prior experience
- Peer annotation: 21 peers plus the baseline
  - New procedure: “peer” review
- Only modified pyramid score (normalized to average # SCUs per model for each pyramid)
Brief Comparison with 2005

- Same characteristics for document clusters
- 4 instead of 7 model summaries
  - 2005: mean of mean SCU weight = 1.9
  - 2006: mean of mean SCU weight = 1.56
- Possibly simpler task (cf. Litowski, DUC 2006)
- Possibly more coherent pyramids
- Improved systems
  - 19/25 (76%) beat the baseline in 2005
  - 17/21 (81%) beat the baseline in 2006
ANOVA Results

- Dependent variable: modified score
- 9 Factors:
  - Peerid ($p \approx 0$)
  - Setid ($p \approx 0$)
  - 5 LingQuality ratings
  - Content responsiveness ($p = 0.0001$)
  - Overall responsiveness (includes readability)
## System Differences (Tukey’s HSD)

<table>
<thead>
<tr>
<th>Peers</th>
<th>&gt; peers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 17, 18, 25, 25 (<em>N=5</em>)</td>
<td>NIL</td>
</tr>
<tr>
<td>22, 29, 32 (<em>N=3</em>)</td>
<td>1</td>
</tr>
<tr>
<td>19, 24, 33 (<em>N=3</em>)</td>
<td>1, 35, 17, 18 (<em>N=4</em>)</td>
</tr>
<tr>
<td>2, 3, 6, 14, 15 (<em>N=5</em>)</td>
<td>1, 35, 17, 18, 25 (<em>N=5</em>)</td>
</tr>
<tr>
<td>28</td>
<td>1, 35, 17, 18, 25, 29 (<em>N=6</em>)</td>
</tr>
<tr>
<td>27</td>
<td>1, 35, 17, 18, 25, 29, 32, 22 (<em>N=8</em>)</td>
</tr>
<tr>
<td>8</td>
<td>1, 35, 17, 18, 25, 29, 32, 22, 14 (<em>N=9</em>)</td>
</tr>
<tr>
<td>10, 23</td>
<td>1, 35, 17, 18, 25, 29, 32, 22, 14, 19, 5, 33, 24, 3, 6, 2, 15 (<em>N=17</em>)</td>
</tr>
</tbody>
</table>
For Illustration: Group Means

<table>
<thead>
<tr>
<th>Peers</th>
<th>Mean modified score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 17, 18, 25, 35 (N=5)</td>
<td>.113 (Δ ~ .06)</td>
</tr>
<tr>
<td>22, 29, 32 (N=3)</td>
<td>.169</td>
</tr>
<tr>
<td>19, 24, 33 (N=3)</td>
<td>.176</td>
</tr>
<tr>
<td>2, 3, 6, 14, 15 (N=5)</td>
<td>.199</td>
</tr>
<tr>
<td>28</td>
<td>.205</td>
</tr>
<tr>
<td>27</td>
<td>.210</td>
</tr>
<tr>
<td>8</td>
<td>.214</td>
</tr>
<tr>
<td>10, 23</td>
<td>.241 (Δ ~ .03)</td>
</tr>
<tr>
<td>Docsets</td>
<td>Mean pyramid score</td>
</tr>
<tr>
<td>-----------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>5</td>
<td>.065 (~0.06)</td>
</tr>
<tr>
<td>1, 3, 8, 15, 47</td>
<td>.133</td>
</tr>
<tr>
<td>50</td>
<td>.135</td>
</tr>
<tr>
<td>45, 30</td>
<td>.158</td>
</tr>
<tr>
<td>28</td>
<td>.164</td>
</tr>
<tr>
<td>16, 17, 20, 29</td>
<td>.172</td>
</tr>
<tr>
<td>27</td>
<td>.197</td>
</tr>
<tr>
<td>14</td>
<td>.229 (~0.03)</td>
</tr>
<tr>
<td>43</td>
<td>.252</td>
</tr>
<tr>
<td>40</td>
<td>.269</td>
</tr>
<tr>
<td>24</td>
<td>.286</td>
</tr>
<tr>
<td>31</td>
<td>.357 (~0.07)</td>
</tr>
</tbody>
</table>
Content Evaluation

- Perfect correlation with mean pyramid score per content level

<table>
<thead>
<tr>
<th>Content Assessment</th>
<th>Mean Pyr Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.12</td>
</tr>
<tr>
<td>2</td>
<td>.17</td>
</tr>
<tr>
<td>3</td>
<td>.19</td>
</tr>
<tr>
<td>4</td>
<td>.21</td>
</tr>
<tr>
<td>5</td>
<td>.22</td>
</tr>
</tbody>
</table>
Comparison with DUC 2005

- Many more significant differences among peers using Tukey
  - 2005: 2 distinct comparison sets
  - 2006: 8 distinct comparison sets

- Better correlation with responsiveness
  - 2 assessor(s) in 2005, $r=.81$; $r=.90$
  - 1 assessor in 2006, $r=1$
Factors Affecting System Scores

- Differences in document set difficulty/coherence
- Pyramid characteristics
  - Mean SCU weight
  - Pyramid size and proportion of weight 1 SCUs
- Score variability
  - 2005: sd = .14
  - 2006: sd = .09
- Better systems
  - 2005 mean system score range: .20 to .06
  - 2006 mean system score range: .24 to .11
Semantics of Pyramids

- More highly weighted SCUs
  - more general
  - less dependent on meaning of other SCUs
Generality of Highly Weighted SCUs

- \( W = 4 \)
  - D0603: *Wetlands help control floods*
  - D0605: *Exercise helps arthritis*

- \( W = 1 \)
  - D0603: *In underdeveloped countries the increase of rice-planting has negative impacts on wetlands*
  - D0605: *Arthroscopic knee surgery appears to reduce pain, for unknown reasons*
Semantic Independence of Highly Weighted SCUs

- $W=4$
  - D0640: *The Kursk sank in the Barents Sea*
  - D0617: *Egypt Air Flight 990 crashed*

- $W=1$
  - D0640: *The escape hatch [of *] was too badly damaged to dock in 7 attempts*
  - D0617: *Tail elevators [of*] were in an uneven position, indicating a possible malfunction*
Impressions/Questions

- Does greater difficulty of a docset correlate with greater specificity/interrelatedness?
  - D0647 is associated with lower mean pyramid scores
  - 9 SCUs of W=4 are all very specific (about sea rescue of Cuban child, Elian Gonzales)
  - 5 of 9 SCUs of W=4 refer to other SCUs
Conclusion

- Systems have improved: DUC roadmap has been successful
- Evaluation document sets have good coverage; but can we begin to characterize document set difficulty?
- Would pyramid scores (intrinsic) correlate with any extrinsic measures?