Pyramid Model of Summary Content

1. Collect model summaries
2. Manual annotation of models for weighted Summary Content Units (SCUs)
   - Mark SCUs: word sequences (possibly discontinuous) that express one idea
3. Match SCUs across summaries: # summaries = SCU weight
4. Iterate over steps a) and b)
5. Generates a pyramid

Properties of Pyramid Method
- No a priori semantic units
- Reliable annotation (Alpha from 0.69 to 0.89 on 10 pyramids)
- Reliable system ranking (Pearson from 0.71 to 0.96 for 16 systems on 8 document sets)
- Existing guidelines used for many studies
- Needs accurate automation

Application to Reading Comprehension
- Archival data from study reported in (Perin et al., 2013)
- 322 community college freshman in remedial classes
- Students read a science text
- Then wrote a summary (varied lengths)
- Reading comprehension score
- Investigators designed a main ideas score
- Represents how many of the text’s main ideas occur in students’ summaries
- Interrater reliability: Pearson correlation = 0.92

Hypothesis: Pyramid Scores Can Substitute for Main Ideas Score
- Models: written by 5 Teachers College masters students
- Targets: 20 of the student summaries
- Pyramid constructed from the 5 models
- Targets scored using normalization 83
- Pearson correlation of pyramid score and main ideas score: 0.85

Correspondence of main ideas and SCUs (60 in pyramid)
- All main ideas occur in pyramid
- Main ideas: SCUs with weight > 2
  - 3 main ideas SCUs with weight 5 (out of 3)
  - 4 main ideas SCUs with weight 4 (out of 7)
  - 6 main ideas SCUs with weight 3 (out of 13)

Automated Pyramid Scores
- Dynamic programming
  - Compare every target ngram to every SCU
  - Compare to SCU label and each contributor
  - Take min, mean, or max score
  - Optimize for the highest unnormalized score

Weighted Matrix Factorization (WMF)
- Latent vector model
  - Assigns a small weight to missing words
  - Suits to semantic comparison of short texts
- 100-dimensional latent vector representation for every ngram in target summaries, every SCU

<table>
<thead>
<tr>
<th>Evaluation Metrics and Results</th>
</tr>
</thead>
</table>

- Use a correlation test to measure how well the automated method produces the same relative ranking
- Use a t test to reject the null hypothesis that the manual and automated score means are the same

<table>
<thead>
<tr>
<th>Summary</th>
<th>Pearson, rank</th>
<th>mean</th>
<th>Diff</th>
<th>T test</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>LVC, max</td>
<td>0.93, 1</td>
<td>49.9</td>
<td>15.65</td>
<td>0.0011</td>
<td></td>
</tr>
<tr>
<td>RO, mean</td>
<td>0.93, 3</td>
<td>48.9</td>
<td>15.60</td>
<td>0.0012</td>
<td></td>
</tr>
<tr>
<td>RO, mean</td>
<td>0.92, 4</td>
<td>47.7</td>
<td>13.45</td>
<td>0.0046</td>
<td></td>
</tr>
<tr>
<td>LVC, max</td>
<td>0.91, 8</td>
<td>52.7</td>
<td>18.50</td>
<td>0.0002</td>
<td></td>
</tr>
<tr>
<td>LVC, min</td>
<td>0.92, 2</td>
<td>37.5</td>
<td>3.30</td>
<td>0.4672</td>
<td></td>
</tr>
</tbody>
</table>

- Compare how well the manual and automated method find the same SCUs
  - There are many ways to achieve the same sum: many combinations of SCUs can get the same numeric score
  - F measure balance how many true SCUs were found (recall) and how many found SCUs were true (precision)

<table>
<thead>
<tr>
<th>SCU 10S</th>
<th>Weight = 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label</td>
<td>Matter is what makes up all objects or substances</td>
</tr>
<tr>
<td>Model 1</td>
<td>Matter is what makes up all objects or substances</td>
</tr>
<tr>
<td>Model 2</td>
<td>Matter is the stuff that all objects and substances in the universe are made of</td>
</tr>
<tr>
<td>Model 3</td>
<td>Matter is identified as being present everywhere and in all substances</td>
</tr>
<tr>
<td>Model 4</td>
<td>Matter is all the objects and substances around us</td>
</tr>
<tr>
<td>Main Idea</td>
<td>All things (objects and substances) in the universe are made of.</td>
</tr>
</tbody>
</table>

SCU 10S Weight = 4

| Distribution of scores for manually identified SCUs given by each similarity for max (a), min (b), mean (c). |

- One idea per sentence
- Each pyramid is a subgraph of the universe
- Pyramid is a collection of ideas
- Pyramid is a representation of ideas

CONCLUSIONS
- In contrast to previous work, good performance on automated scores for individual summaries
  - Promising method to score reading comprehension
  - Accurate, automated content score
  - WMF + cosine gives the best results
  - Use of icdf supports general specification of thresholds across similarity methods

*Center for Computational Learning Systems
†Department of Computer Science
‡Teachers College