Automatic Technique for Linguistic Quality Assessment of Students’ Essays Using Automatic Summarizers

Seemab Latif

School of Computer Science
University of Manchester

18th Nov, 2008
Outline

1. Introduction
   - Story Behind
   - Research
2. Automatic Summarizer
   - Techniques
   - Components
3. Evaluation of Automatic Summaries
   - Experimental Setup
   - Evaluation Outcomes
4. New Finding
   - Verifying My Observation
   - Results
5. An Automated Technique
6. Summary and Conclusion
Conducted an experiment for the evaluation of automatic summarizers
16 annotators participated in this experiment
Experiment was designed to provide a benchmark
It has produced lots of interesting results
Come-up with totally unexpected, un-looked-for new results
Pre-processing the text before document clustering in a way that:

- Computational performance of the algorithm is increased
- Resulting clusters are cohesive and concise

Apply results to the clustering of student’s text answers in an electronic assessment project (ABC) Sargeant et al., 2004, Wood et al., 2006
Research Plan

I. Summarization
II. Summarization → Clustering
III. Summarization → Clustering → Assessment
What is Automatic Summarization?

Automatic text summarization is a process of condensing a source document into a shorter version of text, while keeping the most significant information using a computer program.
Automatic Summarization Techniques

- **Shallow Approach**
  - **Supervised Summarizer**: Supervised summarizer is the one which is dependent on the external input of keywords.
  - **Unsupervised Summarizer**: Unsupervised summarizer is the one which is independent of the external input of keywords. It generates its own keywords by analyzing the source text.

- **Deeper Approach**
Components of Automatic Summarizer

- Pre-processing
  - Case Folding
  - Stemming
  - Stopword Removal
- Keyword Generation
- Sentence Weighting
- Selection
- Production
Components of Automatic Summarizer

- Students’ Answers or Students’ Essays
- Pre-processing
  - Case Folding
  - Stemming
  - Stopword Removal
- Summary Type
- Unsupervised
- Supervised
- Keywords
  - Keywords
  - Summary Type: Supervised or Unsupervised
  - Control the size of the summary
  - Pre-processing
  - Case Folding
  - Stemming
  - Stopword Removal
  - Summary Type
  - Unsupervised
  - Supervised
  - Keywords
  - Summary Type: Supervised or Unsupervised
  - Control the size of the summary
  - Pre-processing
  - Case Folding
  - Stemming
  - Stopword Removal
  - Summary Type
  - Unsupervised
  - Supervised
  - Keywords
- Keyword Generation
  - Identify Keywords in the document
  - Weight sentences based on keywords and heuristics
  - Find the relevant sentences
  - Extract the highest scored sentences
- Sentence Weighting
- Selection
- Production
- Extracted Summary

Seemab Latif  Text Mining and NLP
Keywords Identification

- Students’ Answers - Model Answer and Question
- Students’ Essays - Title or Topic of the essay
- Simple Writing - TF/IDF of Features/Words in the document or Heading of the Document (if present)
Sentence Weighting

- Frequency: count the number of times keywords appear in sentence $S$
- Position: position of the sentence $S$ in the paragraph
- Location: location of the sentence in the document
- Length: calculated by SD and average sentence length.
- Cue Phrases:
  - List 1: weight = 1, e.g. "are analysed", "comparison was made"
  - List 2: weight = 2, e.g. "experimental results demonstrate"
  - List 3: weight = 3, e.g. "my results suggest"
- Stigma Phrases: weight = -1, e.g. "for example", "would like to thank"
- Weight of a sentence is the sum of all individual weights
Selection and Production

Selection phase...
- Identifies parts of a text that can be eliminated
- Removes sentence less than 5 words or greater than 50 words

Production phase...
- Controls the size of the summary
- Writes the selected sentences to summary
Aims for Evaluating Automatic Summarizers

- To evaluate the performance of the summarizers.
- To get computer generated summaries evaluated by humans for qualitative evaluation.
- To have human written summaries as reference summaries for quantitative evaluation.
Experimental Setup

- Two Tasks
  - Creation Task
  - Evaluation Task
- 8 TOEFL Essays
  - 4 for Creation Task
  - 4 for Evaluation Task
  - Why TOEFL Essays?
- 16 Participants
- Summary type: Extractive Supervised and Extractive Unsupervised
- Compression rate: 50%
- Unit for extraction: Sentences
Qualitative Evaluation: Using a Questionnaire

- Coherence: how does the summary read?
- Compactness: how complete is the summary?
- Informativeness: is the content preserved?
- Effectiveness: overall impression of the summary.
Qualitative Evaluation

- Turquoise summary has low scores for human evaluation.
- Teal summary has preserved more information than Turquoise summary.
- Aqua and Cyan summaries have higher scores than the other two.
- Human assigned marks and evaluations are correlated.

<table>
<thead>
<tr>
<th>Essay Name</th>
<th>Human Evaluation</th>
<th>Coherence</th>
<th>Compactness</th>
<th>Informativeness</th>
<th>Effectiveness</th>
<th>Human Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aqua</td>
<td>8.09</td>
<td>8.19</td>
<td>7.94</td>
<td>8.49</td>
<td>7.75</td>
<td>6.0/6.0</td>
</tr>
<tr>
<td>Cyan</td>
<td>7.63</td>
<td>7.93</td>
<td>7.5</td>
<td>7.807</td>
<td>7.28</td>
<td>6.0/6.0</td>
</tr>
<tr>
<td>Teal</td>
<td>6.87</td>
<td>6.37</td>
<td>6.37</td>
<td>7.578</td>
<td>7.19</td>
<td>5.0/6.0</td>
</tr>
<tr>
<td>Turquoise</td>
<td>5.64</td>
<td>6.0</td>
<td>5.93</td>
<td>5.75</td>
<td>5.0</td>
<td>4.0/6.0</td>
</tr>
<tr>
<td>Corr.</td>
<td>0.982</td>
<td>0.950</td>
<td>0.953</td>
<td>0.934</td>
<td>0.915</td>
<td></td>
</tr>
</tbody>
</table>
The experimental results confirmed that the automatic summarizers (developed for this research) have generated summaries that are of good quality, correlate well with human written summaries and obtain a reasonable performance.
It appears that *linguistic quality* of the source text correlates with the degree of *human inconsistency* in content selection in summarization.
Aims for Evaluating My Observation

- To find the correlation between the linguistic quality of source texts and *Human* variations in generating summaries.
- To investigate the same correlation between the linguistic quality of source text and disagreement among different *Automatic* summarizers.
Experimental Setup

- Measure the correlation between the degree of inconsistency among the participants and the linguistic quality of the essays.
- Calculate Degree of Disagreement among Participants (Human and Automatic Summarizers) using ROUGE
- Human assigned marks give the linguistic quality of the essays
What is ROUGE?

- ROUGE (Lin and Hovy 2003) - Recall-Oriented Understudy for Gisting Evaluation
- Automatically evaluates computer generated summaries using unigram co-occurrences
- Does not assume that there is a single gold standard summary
- Standard objective evaluation measure for the Document Understanding Conference (http://duc.nist.gov/)
ROUGE Scores

- **ROUGE-1** is unigram overlap between system (computer generated) and reference (human generated) summaries.
- **ROUGE-2** is bigram overlap between system (computer generated) and reference (human generated) summaries.
- **ROUGE-L** is the longest common word subsequence between system and reference summaries.
- **ROUGE-W** is same as ROUGE-L but gives high weighs to consecutive occurring words.
Degree of Disagreement among Humans

- ROUGE for evaluation
- 16 human generated summaries
- select one summary as system summary
- other 15 as reference summaries
- repeat this process for each summary
- calculate average values over all 4 documents
Blue and Green summaries have higher participant agreement.

Red and Yellow summaries have high inconsistency.

ROUGE and Human assigned scores are highly correlated.
Degree of Disagreement among Automatic Summarizers

- 9 automatic summarizers
- 12 essays- organised in 3 datasets
- select one summary as system summary
- other 8 as reference summaries
- repeat this process for each summary
- calculate average values over all 3 datasets
Degree of Disagreement among Automatic Summarizers: Findings

<table>
<thead>
<tr>
<th>Dataset I</th>
<th>ROUGE</th>
<th>Human Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Essay 1</td>
<td>61.96</td>
<td>54.09</td>
</tr>
<tr>
<td>Essay 2</td>
<td>67.65</td>
<td>60.97</td>
</tr>
<tr>
<td>Essay 3</td>
<td>69.54</td>
<td>59.56</td>
</tr>
<tr>
<td>Essay 4</td>
<td>74.02</td>
<td>61.35</td>
</tr>
<tr>
<td>Corr.</td>
<td>0.872</td>
<td>0.592</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset II</th>
<th>ROUGE</th>
<th>Human Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Essay 1</td>
<td>57.38</td>
<td>46.79</td>
</tr>
<tr>
<td>Essay 2</td>
<td>57.27</td>
<td>49.05</td>
</tr>
<tr>
<td>Essay 3</td>
<td>72.06</td>
<td>60.62</td>
</tr>
<tr>
<td>Essay 4</td>
<td>74.64</td>
<td>65.99</td>
</tr>
<tr>
<td>Corr.</td>
<td>0.810</td>
<td>0.895</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset III</th>
<th>ROUGE</th>
<th>Human Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Essay 1</td>
<td>63.93</td>
<td>54.37</td>
</tr>
<tr>
<td>Essay 2</td>
<td>64.69</td>
<td>55.08</td>
</tr>
<tr>
<td>Essay 3</td>
<td>66.63</td>
<td>55.85</td>
</tr>
<tr>
<td>Essay 4</td>
<td>74.57</td>
<td>60.93</td>
</tr>
<tr>
<td>Corr.</td>
<td>0.992</td>
<td>0.993</td>
</tr>
</tbody>
</table>
Assessing the linguistic quality of the essay by summarizing the source essay using a number of automatic summarizers and then analyzing the degree of inconsistency among the automatic summarizers.
Applications of New Technique

It can be used:

- to automatically divide the students essays into broader bands of quality
- to give initial feedback to the students about the quality of the essay they have written
- in e-learning environment
Summary

- Low quality essays have low quality summaries.
- ROUGE evaluation of students’ essays has a strong positive correlation with the human evaluation of those essays.
- By analyzing the correlations between ROUGE evaluations and human evaluations of the students’ essays, we get an automated way of assessing the linguistic quality of students’ essays.
Conclusion

- Devised a novel and automated method...
- Through experimentation and analysis...
- Lot of interest these days...
Acknowledgements

- I would like to acknowledge the participation of students who took part in this experiment.
- I would like to gratitude my supervisors Mary McGee Wood and Goran Nenadic for their support, encouragement, useful feedback and reviewing my work.

THANK YOU