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

Story behind...

- 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

About Overall Research

- 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 $\rightarrow$ Clustering
III. Summarization $\rightarrow$ Clustering $\rightarrow$ 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.

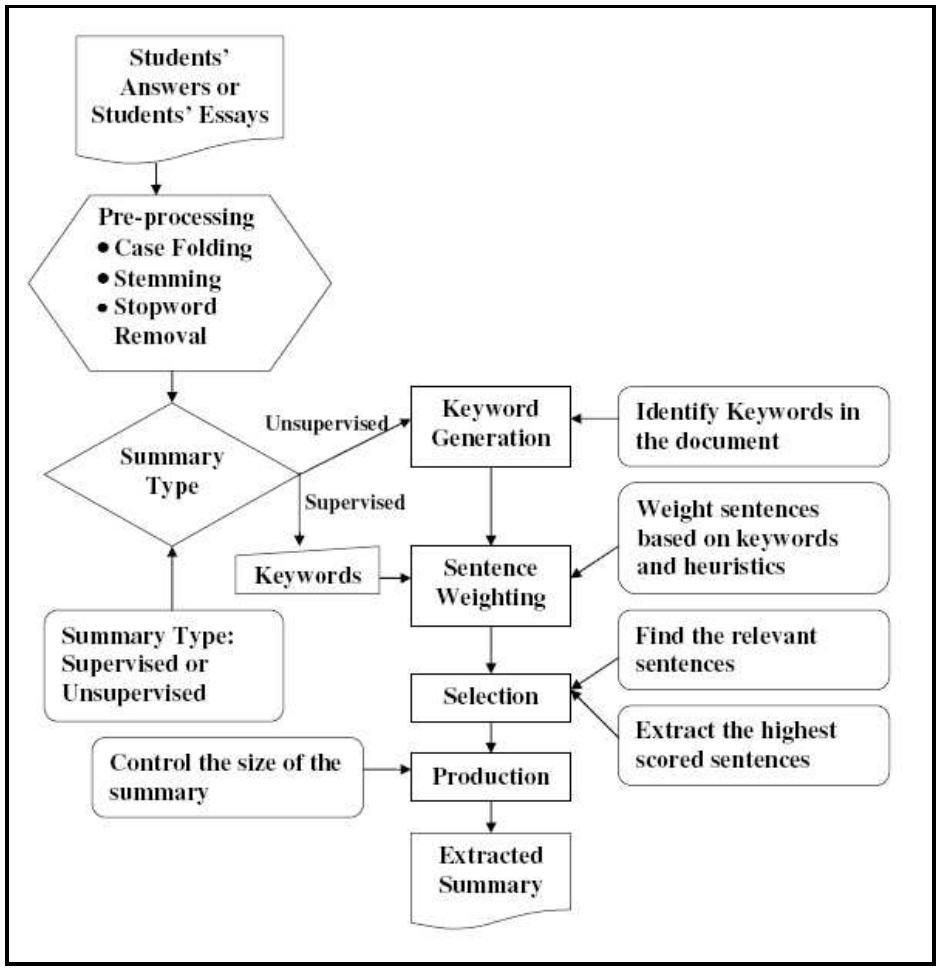
Automatic Summarization Techniques

- 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



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:
  - List1: weight = 1, e.g. "are analysed", "comparison was made"
  - List2: weight = 2, e.g. "experimental results demonstrate"
  - List3: 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 then Turquoise summary.
- Aqua and Cyan summaries have higher scores then the other two.
- Human assigned marks and evaluations are correlated.

Essay Name	Human Evaluation	Coherence	Compactness	Informativeness	Effectiveness	Human Marks
Aqua	8.09	8.19	7.94	8.49	7.75	6.0/6.0
Cyan	7.63	7.93	7.5	7.807	7.28	6.0/6.0
Teal	6.87	6.37	6.37	7.578	7.19	5.0/6.0
Turquoise	5.64	6.0	5.93	5.75	5.0	4.0/6.0
Corr.	0.982	0.950	0.953	0.934	0.915	

Evaluation Results

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.

Interesting Observation

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

Degree of Disagreement among Humans: Findings

- Blue and Green summaries have higher participant agreement.
- Red and Yellow summaries have high inconsistency.
- ROUGE and Human assigned scores are highly correlated.

Text *Name*	ROUGE				Human *Assigned* Scores
	1	*2*	*L*	*w-1.2*	
Red	67.26	52.11	65.68	21.36	5.0/6.0
Blue	73.95	64.47	73.71	26.62	6.0/6.0
Green	73.20	61.27	72.43	24.18	6.0/6.0
Yellow	67.56	59.19	67.26	21.25	4.5/6.0
Corr.	0.948	0.615	0.895	0.891	

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

Dataset *I*	ROUGE				Human *Scores*
	1	*2*	*L*	*w-1.2*	
Essay 1	61.96	54.09	61.17	28.11	1.5/6.0
Essay 2	67.65	60.97	66.81	28.59	2.0/6.0
Essay 3	69.54	59.56	68.64	27.96	2.5/6.0
Essay 4	74.02	61.35	73.44	29.21	6.0/6.0
Corr.	0.872	0.592	0.88	0.84	

Dataset *II*	ROUGE				Human *Scores*
	1	*2*	*L*	*w-1.2*	
Essay 1	57.38	46.79	56.93	28.48	0.0/6.0
Essay 2	57.27	49.05	57.03	32.20	2.0/6.0
Essay 3	72.06	60.62	71.98	36.83	2.5/6.0
Essay 4	74.64	65.99	73.32	39.85	6.0/6.0
Corr.	0.810	0.895	0.786	0.932	

Dataset *III*	ROUGE				Human *Scores*
	1	*2*	*L*	*w-1.2*	
Essay 1	63.93	54.37	63.30	27.83	2.0/6.0
Essay 2	64.69	55.08	64.07	32.51	2.0/6.0
Essay 3	66.63	55.85	66.42	31.94	2.5/6.0
Essay 4	74.57	60.93	74.28	37.92	6.0/6.0
Corr.	0.992	0.993	0.988	0.882	

An Automated Technique

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.

Publications

- Seemab Latif and Mary McGee Wood, "Text Pre-Processing for Document Clustering", In Proceedings of Natural Language and Information Systems (NLIS), Lecture Notes in Computer Science, June 2008, pp 358-359.
- Seemab Latif, "Text Pre-Processing for Document Clustering", In Proceedings of Computational Linguistics UK (CLUK) held at University of Oxford Computing Laboratory, March 2008.

THANK YOU