Essay grading by machine

John Broida
Psychology Department
University of Southern Maine
I teach large sections, so cannot grade essay exams.

• With 100 students or more, it is simply impossible. So writing was dropped from my introductory classes.
• But, I value writing – want it in those classes.
• I heard about essay grading programs a few years ago, and tried one in the past year.
  • I am thrilled.
  • So are others: http://www.nytimes.com/2013/04/05/science/new-test-for-computers-grading-essays-at-college-level.html?pagewanted=all&_r=0
• One advantage: not having to deal with hand written papers.
• Another advantage – immediate feedback.
• But perhaps the most important benefit is that students can learn what their mistakes are, and how to correct them
The hard part is the question: It must elicit a specific set of key terms.

- Ideally there should be a single answer to the question
  - Consider the difference between asking
    - what happened on this date 10 years ago (multiple answers)
    - what are the 5 layers of the eye that light passes through on its way to the retina, and what does each of those layers do.

- Once you have a good question, the process is easy. Ask the question of a couple of experts, and several hundred students. Then let the machine go to work.
  - Most programs have the student work graded by machine, which then looks at its routine in light of the grade given by the professor
  - The machine looks for relationships between words, and may also check grammar and spelling. In the ideal, it will check for plagiarism.
These programs look at how words are related – in distance

- Use latent semantic analysis, which assumes that words that are close in meaning will occur in similar pieces of text.
- Are they used together or not, how close one word appears in relation to other words.
- First the researchers do a word relatedness matrix by asking subjects about the degree of relationship between them, 1 is high related (hot, cold) versus 9 not related (and, the).
- They then do a path analysis to determine which combinations are most frequently rated as close versus distant.
- This information is then checked by analyzing an expert answer to the question – does that individual score high using the system. May also relate to a bad answer, as judged by one or more experts.
There may be multiple relationships – for example, 1 in each part of the correct answer

• The systems, in effect, use both quantitative and qualitative systems to evaluate writing.

• Quantitative look at size (how many terms are used in comparison to expert answers) and strength (the number of links between concepts in comparison to those made by experts).

• Qualitative look at spoke (are all concepts mentioned), chain (linear sequence of understanding/concepts and net (the integration and hierarchy of key terms which indicates depth of understanding).
There are two approaches to the grading of essays

• Sentence Aggregate Approach.
  • A two dimensional approach to the links between words.
  • It really does not work as well as the second approach -

• Linear Aggregate Approach.
  • A more traditional approach based on sentence structure.
  • Allows one to see the effects of repeated use of terms.

• In what follows, I am borrowing heavily (stealing) from a presentation by Roy B. Clariana, *Pennsylvania State University,*
Text analysis using the sentence aggregate approach

*Humanists believed that* job satisfaction *was related to* productivity. *They found that if* employees *were given more* freedom and power *in* their jobs, *then they* produced *more.*

Here is the sentence aggregate run shown in black.

*Historical basis...an excel spreadsheet to do sentence aggregates*
Text analysis using the linear aggregate approach on an edited version of the text

*Humanists* believed that *job satisfaction* was related to *productivity*. They found that if *employees* were given more *freedom and power* in their *jobs*, then they *produced more*.

Here is the linear aggregate run shown in black, so far so good, but what happens on the 2\textsuperscript{nd} occurrence of ‘job’
Humanists believed that job satisfaction was related to productivity. They found that if employees were given more freedom and power in their jobs, then they produced more.

Now the word job appears twice, so the machine has two arrows to the term.
Humanists believed that job satisfaction was related to productivity. They found that if employees were given more freedom and power in their jobs, then they produced more.

The force-directed graph begins to fold back on itself because job now refers to two ideas.
Humanists believed that job satisfaction was related to productivity. They found that if employees were given more freedom and power in their jobs, then they produced more.
Do they work?

• I have experience with only one program but, based on that, I would say the answer to this question is yes.

• The quality of writing improves with each draft of the answer.

• The student reaction has been favorable – despite the $25 or so that they paid (I now have a deal that will cost them $3 for the semester)

• They like immediate feedback, and the chance to fix their answers.
Now you can see if these things work...

• One program is readily available, and we can fake it with two others
  • SAGrader.com – the one that I use.
    • Go to http://SAGrader.com and create a student account.
    • Use the access code FONMFAJKHW
    • The thing will ask for an ID number – Pick one that you will remember.
    • This is a student account. I can show you what the instructor sees.

• The Pearson Publishing Version
  • Go to http://mypsychlab.com

• The Cengage Publishing Version
  • I hope to have access soon. Let me know if you are interested and I will provide the information when I get it.