The Evolution of Automated Writing Evaluation

*Is the writing on the ... “page”?*

Nitin Madnani / Distinguished Research Engineer / ETS
Automated Evaluation of Writing – 50 Years & Counting
Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2020

Joint work with Beata Beigman Klebanov
Automated essay grading system with 32 features combined via linear regression; $r=0.65$ with average human score on 276 essays by high school students.

Reduce load on teachers and facilitate fast turnaround for feedback.

Automated Writing Evaluation (AWE) is *worthwhile* and *achievable* for a *specific goal*: not a “master” analysis of the writing *a la* a human reader but an *imitation* that produces a *correlated result*.

Thoughtful discussion of various AWE opportunities & challenges.
Report Card

Are we still on the same ... “page”?

- Notable Achievements
- Needs Improvement
- Off the “Page”
Notable Achievements

• AWE systems today can score in agreement with the average human rater, in many contexts:
  • ACT Next’s CRASE+®
  • ETS’s eRater®
  • Measurement Inc’s Project Essay Grade®
  • Pearson’s Intelligent Essay Assessor®
  • Vantage Learning’s Intellimetric®
• Automated and human scores are often used together (weighted combination, check score).
Needs Improvement: Originality

What about the gifted student who is off-beat and original? Won’t he be overlooked by the computer?

• Page: once we can measure originality objectively, we can add it as a feature to the scoring system.

• Existing work on measuring characteristics of outstanding writing.

• Aspects of language use that are often considered original have been studied in the context of essay evaluation.


Needs Improvement : Gaming

Won’t this grading system be easy to con? Can’t the shrewd student just put in proxies which will get a good grade?

- Page: the grading program may come to consider so many variables that the best way to “con” it is to write well.

- Generally handled using small accompanying programs (advisories) for spurious lengthening, varying sentence structure, replacing words with sophisticated variants, plagiarism, unnecessary “shell” language, etc.

- Higher stakes engender a never-ending battle of wits.


Needs Improvement: Content

We are talking awfully casually about grading subject matter like history … Aren’t we supposed to see what the students are saying makes sense … ?

- Adjust the AWE system to attend to details of genre and task
  - Appropriate use of specific source materials
  - Quality of specific narrative, reflective, and argumentative elements
- Content scoring is now a parallel line of research
  - Dedicated scoring model for every question with fluency deemed secondary


So far, the work looks like grading, not correcting. Isn’t the need much greater for correction and comment?

- Language conventions & grammar feedback is incorporated into text editors
- Many tools strive to provide more complex feedback
  - Discourse Structure (Criterion®)
  - Tone (WritingAssistant®)
  - Thesis Relevance (Writing Pal®)
  - Topic Development (Writing Mentor®)
- Research on the effectiveness of automated feedback is inconclusive
Off the “Page” : Multilinguality

- Methods developed for one language/dialect may not generalize
- Active area of research for multiple languages
  - Arabic
  - Chinese
  - Danish
  - Finnish
  - French
  - German
  - Japanese
  - Norwegian
  - Portuguese
  - Swedish
  - Thai

Off the “Page” : Standardized Testing

• Ensure that scores are **valid** (measure intended skills)

• Ensure that scores are **defensible** (clear post-hoc explanation)

• Ensure that scores are **fair to all test-takers** (no undue advantage of race, ethnicity, gender, age, socio-economic status, linguistic/cultural background, test characteristics)

• Ensure that scoring system is **scalable, reliable, and flexible** to support large-scale use
Off the “Page”: Pervasive Technology

- Page’s thought experiment: classroom-first (only?) use for AWE
- Reality: carry a powerful computer (and, by extension, AWE systems) in your pocket and use it (almost) anytime, anywhere, for anything
- A writing aid meant to help a student construct better arguments could also end up being used by a lawyer to draft his closing argument
- How does one evaluate a technology without knowing what it could be used for?
A Taxonomy of AWE Use Cases
Support Consequential Decision-Making

• About the writer or a related entity based on the written product
• The emphasis is on providing explainable, fair, and valid scores
• Examples:
  • Standardized assessments for higher-ed admissions
  • Licensure exams for professional certifications
  • Job applications that require a writing sample
  • Course placement decisions
Create a Better “Written Product”

- Focus is on the **actual piece of writing** and its **real-world impact**
- Distinction of **human- vs machine-produced** is **irrelevant**
- Machine-augmented Human > Human
- Examples
  - More engaging blog post that **increases click-through rates**
  - More impactful advertising copy that **increases sales**
  - More professional-looking email that **increases survey participation**
3 Help Writers Improve Their Skills

• Feedback (or scores) designed to help users imbibe writing skills

• First human-only draft of next essay > first draft of current essay

• Difficult to give examples; controlled measurement of skills is hard!

• Not necessarily mutually exclusive with the other 2 use case types …
AWE Use Types Can Overlap

• Example: allowing spell-correction software on a standardized test

• Human augmentation + consequential decision-making

• Manually-vetted spell correction significantly improves scores assigned by trained human raters to weaker writers

• Spell-correction software is less accurate for essays by weaker writers

• Overlapping use cases require careful examination of priorities, e.g., validity and fairness in this case


AWE Use Types Can Conflict

- Consistent, pervasive human augmentation may impact skill acquisition
- Finding & fixing identified errors ≠ skill-building
- Could spelling end up the way of “long division” and “complex paper-and-pencil computations”?
- “Use it or lose it”?


Summary

• Page’s 1966 paper with a proof-of-concept demonstration and an outline of associated challenges (mostly) stands the test of time.

• Page imagined AWE as mainly serving an English teacher, not standardized testing or a PR executive running Grammarly on a press release – although he foresaw some of the challenges.

• "Almost universal" AWE can provide value in different contexts - decision-making, human augmentation, and skill improvement.

• As NLP practitioners, we can help by using context-driven design & evaluation and by engaging the right partners.