White Paper: PEG Changes

Michael B. Bunch, Thomas Davis, Ann Hayes, Derek Justice, Julie St. John
Measurement Incorporated
July 2017

Introduction

At the close of World War II, the world learned of the heroic codebreaking efforts of Alan Turing and his colleagues in British intelligence. The theory behind the accomplishments of the Turing group gave rise to the new field of artificial intelligence (AI). It was only a matter of time before someone would apply the theory and concepts of AI to a task of vital importance to educators: the grading of student written essays by a machine.

Ellis Batten Page (1924–2005) is widely acknowledged as the father of automated essay scoring. Page (1966) reported on an early effort to understand how human beings graded student essays and to translate the process into a computer program. That program, Project Essay Grade, or PEG®, was designed to score student essays using mainframe computers in the 1960s.

As a result of Page's work, two new terms entered the lexicon: *trin* and *prox*. A trin is an intrinsic characteristic of writing, such as diction or style. A prox is a quantifiable approximation of that intrinsic characteristic. These terms have since been replaced by "features," and there is no practical distinction between intrinsic and objectified features. "Artificial intelligence," at least in this context, has been replaced by "automated essay scoring or "automated essay evaluation."

The initial PEG work focused on essays written by 276 high school students and graded by four English teachers. Those essays yielded 31 proxes (assigned by PEG in accordance with rules devised by Page) used as predictors of scores assigned by teachers. Page and his colleagues calculated the correlation between a weighted composite of the 31 proxes and the scores assigned by teachers. The resulting multiple R was .71. When one considers the fact that the correlation between scores assigned by two English teachers is not much higher, these results were quite remarkable.

Page applied the tools available to him as an English teacher: a deep understanding of the intrinsic qualities of good writing (*trins*) and the ability to translate those qualities into objective units (*proxes*). He then applied the tools available to him as a psychometrician: multiple regression and the ability to interpret its results. In doing so, he created the field of automated essay scoring (AES).

As AES has matured over the past 50 years, and as trins and proxes have given way to features, the goal of programs like PEG has been to improve predictability of human-rendered scores. As

multiple R has also given way to more sophisticated metrics (e.g., quadratic weighted kappa, or *QWK*), that goal has evolved into increasing the size of *QWK*, specifically, achieving a *QWK* for AES equal to or greater than a *QWK* for human-rendered scores.

That goal was officially reached in 2012. Documenting the first Automated Scoring Assessment Prize (ASAP) competition, Morgan, Shermis, Van Deventer, & Vander Ark (undated) reported that five vendors' automated essay scoring programs had surpassed human readers in score stability. Since that time, the race to increase *QWK*, even incrementally, has continued. Larger and larger values of *QWK* have been achieved, primarily by the addition of features. At some point, however, the number of features grows so large that interpreting results becomes a challenge.

Continuous Improvement for PEG and Writing Assessment

Measurement Incorporated purchased PEG from Dr. Page in 2003. Since that time, we have updated and modified the software on a regular basis. The 2012 ASAP competition (in which MI/PEG took first place) was an important milestone in the history of PEG, but it was not the only one. Improvement continues. Specifically, as the field of writing assessment moves forward, as the definition of good writing evolves, and as we refine computational procedures, we will modify PEG to provide more reliable, valid scores.

Recent improvements and rationale. PEG is now presented to users in two forms. First, the existing (and recently improved) Peg Web Service, the real-time, formative AI provides prompt-generic scoring and feedback to the students and teachers using Measurement Incorporated's Writing Sites. Second, PEG has now also been made available for prompt-specific, batch-based scoring to answer seasonal demand for large sets of summative scores. PEG's increased availability has increased demand for new functionality, some of which has debuted in the formative service, the summative service, or both. These include the definition of additional features, the expansion of targeted feedback, the creation of an optional rules-based alert language scanner, the introduction of a new method for QWK optimization, and the necessary infrastructure to support anticipated upcoming functionality such as prompt-specific plagiarism detection and improvement of AI model interpretability.

Improvement in formative tools and applicability. Although the transition from 31 trins and proxes to over 300 features has improved PEG's accuracy in scoring, an unintended consequence has been a decrease in the ready explicability of scores. In the formative context, PEG is used to score student writing and provide targeted feedback for improving the essay. The primary objective is to improve the student's writing ability. As such, the feedback generation should be tightly coupled with the scoring engine, so that if a student earnestly follows the suggestions provided, he/she can expect to see an improvement in the score of the next submitted revision.

Assuming the feedback is clear enough, following it should effect changes in certain features that ultimately lead to score changes via the model. Traditional, black-box AI models make

feedback/score coupling difficult because the features enter the models in very complex ways. Indeed, there is no constraint present in traditional models that would ensure a score improvement if some positive feature extracted from the writing is increased.

2017 modifications and rationale. New PEG models for the August 2017 release have been developed to address this problem. They are designed from the formative perspective by selecting a smaller set of instructionally meaningful features around which clear feedback text can be written. The models are explicitly constrained such that if a certain feature that should positively affect the score (i.e., a feature representative of good writing) is increased (due to prompting from the feedback), then the score will necessarily increase. The score is also guaranteed to increase if a feature that should negatively affect the score is decreased.

A student may, for example, receive feedback suggesting that using transitional words will improve her/his essay and that correcting misspelled words will make the essay easier to understand. An increase in the use of transitional words will improve the essay by more closely tying together ideas, resulting in a higher score in the Development of Ideas and Organization traits, while reducing the number of misspelled words results in a higher score in Conventions.

Plans are in place to release an update that will further increase the impact of applied feedback. With this update, feedback will be selected based upon which features are expected to give the greatest score increase of the student's current revision. The student should see an upward trajectory across revisions if the feedback is followed, since anomalies wherein the score drops even though features tied to the feedback are properly adjusted become mathematically impossible with the new models. Also, the contributions to the score for different aspects of writing are saturated, so if the student wants to push the score ever higher, he/she is forced to heed feedback to improve across all aspects of writing.

Once the writer has attended to feedback that increases Development of Ideas, for example, and Organization scores such that those aspects are saturated, the model will provide feedback from another area such as Word Choice or Sentence Fluency. In addition to making the students better writers (which is the purpose of formative tools after all), the new models carry the additional benefits of producing fewer unexpected scoring oddities (since they are explicitly constrained), and being easily interpretable by the teachers (since they are derived from instructionally meaningful features tied directly to the revision feedback).

One other change in the writing sites models for the August 2017 release is the replacement of Sentence Structure with Sentence Fluency, which was the original fifth trait in the Six Traits of Writing. Sentence Structure had included scoring and feedback on parts of grammar such as subject/verb agreement as well as fragments and run-ons. Changing this trait to Sentence Fluency clarifies its purpose and better distinguishes it from Conventions, which now includes all grammar and usage.

What These Changes Mean for You

These changes, increased transparency and accuracy of the new models, the ordering of targeted feedback to most impact scores, and the change from Sentence Structure to Sentence Fluency are designed to increase students' and teachers' understanding of how best to revise essays and increase writing skills.

MI continues to monitor advancements in the automated essay scoring field while searching for ways to make PEG as effective as possible in helping students learn to write. As a result, PEG will be ever-evolving.

References

Morgan, J., Shermis, M. D., Van Deventer, L. & Vander Ark, T. (undated). *Automated Student Assessment Prize: Phase 1 & Phase 2: A Case Study to Promote Focused Innovation in Student Writing Assessment*. Retrieved 9/1/14 from http://gettingsmart.com/wp-content/uploads/2013/02/ASAP-Case-Study-FINAL.pdf

Page, E. B. (1966). The imminence of…grading essays by computer. *Phi Delta Kappan*, 47 (2). 238-243).