Book Review

Automated Essay Scoring

Beata Beigman Klebanov and Nitin Madnani
(Educational Testing Service)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 52), 2022, xx+294 pp; paperback, ISBN 9781636392226; ebook, ISBN 9781636392233; hardcover, ISBN 9781636392240;
doi:10.2200/S01121ED1V01Y202108HLT052

Reviewed by
Anaïs Tack
KU Leuven

Automated essay scoring is concerned with the development of language technologies that make it possible for an essay—or any piece of writing for that matter—to be evaluated or scored by a computer. These technologies find their utility primarily in the context of educational measurement, where they serve a dual purpose. On the one hand, they provide crucial support to educators and institutions, facilitating the assessment of students' writing skills and content knowledge. A good example is the TOEFL iBT, the Internet-based Test of English as a Foreign Language, administered by the Educational Testing Service and widely adopted by institutions worldwide. On the other hand, these technologies benefit writers themselves, including students, by offering a platform to assess and enhance their writing skills. One illustrative tool for this purpose is the Write & Improve software developed at the University of Cambridge.

The field of automated essay scoring emerged in the pioneering era of artificial intelligence and computational linguistics, with the inception of Ellis Page's *Project Essay Grade* system at the University of Connecticut in 1964, and the subsequent publication of his seminal article "The Imminence of…Grading Essays by Computer" in 1966. Page's automated scoring system is seen by the authors of this book as one of the first concrete applications of natural language processing, after the Audrey system for speech recognition and the Georgetown–IBM demonstration of machine translation. But just like speech recognition and machine translation, the industrialization of automated essay scoring mostly gained momentum in the 1990s. This decade saw an increase in richly annotated language data and corpora, which enabled the use of statistical and supervised machine learning in developing essay-scoring systems. Pearson's Intelligent Essay Assessor, released in 1998, is a prominent example of this evolution. A year later, in 1999, the e-rater scoring engine was launched by the Educational Testing Service, commonly abbreviated as ETS, a private organization overseeing several standardized tests and high-stakes examinations such as TOEFL and GRE.

This book is written by Beata Beigman Klebanov and Nitin Madnani, two researchers at ETS with many years of research experience and numerous peer-reviewed

https://doi.org/10.1162/coli_r_00513

publications in the field of automated essay scoring. Their book provides a concise yet indispensable introduction to the field. After this short introduction, the eager reader is invited to take a deep dive into the scientific literature, encompassing slightly over half of the book's content, approximately 115 pages. The literature review primarily caters to computational linguists and NLP practitioners, as it delves comprehensively into diverse machine learning models—ranging from linear regression to artificial neural networks—and intricate linguistic features. These features include general indicators of writing quality, such as text organization, coherence, and grammaticality, as well as genre-specific features, such as argument structure in argumentative essays. Furthermore, the reader not only gains insight into the extensive research carried out at ETS throughout the years but also delves into their technical expertise through a series of guided experiments with RSMTool—an open-source tool developed by the second author and colleagues. In addition, the authors also provide insight into a scalable and production-ready computer architecture used to build ETS products such as c-rater, e-rater, Language Muse, and Writing Mentor.

The book is divided into five parts counting 13 chapters in total. The first part contains an introductory chapter in which the authors introduce the reader to Page's aforementioned seminal 1966 article. Chapter 1 enumerates several arguments made by Page in favor of automated essay scoring, including its educational need, computational feasibility, high quality, and low cost. The chapter then sets forth four challenges associated with automated essay scoring. Two among them are the evaluation of original and source-based writing. Original writing, which reflects an author's unique voice and thus stands out from existing and conventional works, poses a challenge because its originality could be overlooked or even under-evaluated by a computer. Source-based writing, a form of writing that reviews main points from external sources, presents a different problem because the assessment focuses on the correctness of content rather than measuring writing skills and essay quality. Another challenge is avoiding potential gaming strategies that test takers may employ to inflate their scores. A final challenge is automated feedback, as the effectiveness of providing linguistic and stylistic commentary on written work is not always proven.

The second part contains two chapters covering a series of guided experiments and a set of best practices when building an automated essay scoring system. Chapter 2 provides a step-by-step guide for building an automated scoring system with supervised machine learning. First, the reader learns the usual engineering setup, including the use of a standard dataset (viz., the Automated Student Assessment Prize competition), an interpretable machine learning model (viz., linear regression), and a set of basic features derived from a scoring rubric. The authors also introduce the reader to their RSMTool. Then, the reader is guided through a series of ten experiments illustrating the incremental development (experiments 1–5) and evaluation (experiments 6–10) of the machine learning model. In each experiment, the reader is taught an important lesson such as feature fairness. For each experiment, the authors refer the reader to a report (i.e., correction key) available online.

Chapter 3 provides some best practices for building an automated essay scoring system. First, the authors identify potentially conflicting perspectives between NLP developers and other stakeholders, including end users, subject matter experts, and business units. The authors then describe three natural use cases for automatic essay scoring, where it is added to a pre-existing assessment, developed concurrently with a new assessment, or implemented in a classroom. The authors provide some practical considerations and concrete actions that are well thought out and will appeal to many different readers.

The third part contains five chapters that provide an up-to-date overview of the scientific literature and a more detailed account of the concepts introduced in the previous chapters. Chapter 4 describes various statistical models, including linear regression, latent semantic analysis, support vector machines, random forests, ensemble methods, and neural networks. For each of these models, the authors describe their mathematical underpinnings and explain how they can be used to score an essay. The chapter pays special attention to recent deep-learning architectures for automated essay scoring.

Chapters 5 and 6 describe computational features that capture various aspects of the writing construct and the scoring rubric. Chapter 5 deals with general features. The features are organized into three classes: discourse features aiming to capture essay organization, development, and coherence; content features related to vocabulary use and topicality; and conventional features based on grammatical error detection. Chapter 6 then gives an in-depth overview of computational features that pertain to specific writing genres. The chapter is focused on four genres: argumentative writing, narrative writing, source-based writing, and reflective writing. Argumentative writing involves defending a particular position on a given topic (e.g., why technology should be integrated into education), presenting several claims as to why this position is valid, and supporting these claims with premises and evidence. Narrative writing involves telling a story that describes, for example, the historical evolution of technology in education. Source-based writing involves summarizing and comparing key points from external sources to compose an informed essay, which, for example, reviews the effectiveness of integrating technology into education based on scholarly sources. Finally, reflective writing involves examining personal experiences, such as teachers describing their experiences integrating technology into the classroom and reflecting on important lessons learned from this experience. This detailed overview of different writing genres also interestingly introduces the reader to research from related fields, such as computational argumentation and text summarization.

Chapters 7 and 8 address two concerns in setting up real-world applications of automated essay scoring. Chapter 7 deals with the issues of reliability, scalability, and flexibility when deploying a scoring system at a large scale. The chapter describes and illustrates an Apache Storm architecture implemented in several ETS systems. Chapter 8 is about evaluating construct validity and fairness. When deploying an essay scoring system for high-stakes testing, fundamental issues arise when the system fails to measure the construct, assigns scores influenced by factors irrelevant to the measured construct, or is biased towards specific personal characteristics in the population of test takers. If an essay scoring system overlooks important features or favors particular writing styles or cultural representations, it undermines the validity and fairness of high-stakes assessments.

The fourth part contains four chapters examining some broader challenges introduced in the first chapter and which remain to be solved for automated scoring. Chapter 9 deals with the automated generation of useful feedback on writing. The authors review several existing feedback systems and discuss how to define and evaluate the usefulness of feedback. Chapter 10 focuses on evaluating essay content, which is separate from assessing essay quality. Content scoring emphasizes measuring test-takers' content knowledge, prioritizing these elements over writing skills. This evaluation can adopt either a reference-based or a response-based approach. Reference-based scoring involves comparing responses to a set of predefined reference answers, while response-based scoring independently assesses the response content. The chapter primarily explores the latter, investigating computational features and models tailored for this approach. Chapter 11 deals with another task related to but different from essay scoring,

namely the automated scoring of spontaneous speech. After a brief account of the challenges with automated speech recognition, the authors review three sets of features for speech scoring: the delivery and fluency of spontaneous speech, vocabulary and grammar use, and topic development. The authors also contrast features relevant for scoring speech with those relevant for scoring writing. Lastly, chapter 12 examines several gaming strategies test-takers could use to fool the automated scoring system into giving a higher score. The authors review four types of strategies: the unnecessary use of shell language, the artificial generation of essays, the submission of off-topic responses, and the use of canned responses or plagiarized essays.

The fifth and final part of the book contains a concluding chapter. The authors revisit the desiderata put forth by Ellis Page in his 1966 publication and summarize the overall achievements and remaining challenges in this respect. In addition, the authors discuss other challenging aspects that Page did not envision, such as the present-day ubiquity of technology, dealing with multiple languages, and setting up high-stakes tests that are valid, defensible, and fair.

In sum, the book offers an excellent introduction to and deepening of the field of automated essay scoring. The book is well-structured and easy to read. Throughout the book, the authors provide thoughtful insights and practical advice based on their many years of experience at ETS. Compared to other books on the subject, the book offers a valuable combination of practical lessons and scientific deepening. By the end of the book, the reader has acquired a broad knowledge of the possibilities, challenges, and practical concerns involved with the automated scoring of student writing.

Anaïs Tack is a postdoctoral researcher in language technology for smart education at itec, an imec research group at KU Leuven; she is also a visiting lecturer in natural language processing at UCLouvain. She has worked on the computational modeling of linguistic complexity in texts for automated writing evaluation and difficulty prediction. Her e-mail address is anais.tack@kuleuven.be.