Educational Data Mining in Writing and Literacy Instruction

Collin Lynch
NC State University
cflynch@ncsu.edu

Paul Deane
ETS
PDeane@ets.org

Piotr Mitros
ETS
pmitros@ets.org

Zhikai Gao
NC State University
zgao9@ncsu.edu

Damilola Babalola
NC State University
djbabalo@ncsu.edu

ABSTRACT

The emergence and application of new technology, such as generative AI, are bringing new challenges and opportunities for educational researchers and educators. For writing and literacy education, we are facing challenges like LLM-generated text detection and students' data privacy. Still, we also could benefit from automatic essay scoring, personalized instruction, and feedback generation. How we address such challenges and utilize these opportunities are essential questions for our EDM community to foster future writing education. In this workshop, we will discuss the state-of-the-art research and application of writing and literacy education. Moreover, during a tutorial session, we will introduce a prototype platform we are currently developing to support ethical students' writing and learning data management. We believe both the paper presentation and the tutorial session in our workshop could benefit us and the participants through learning opportunities and potential collaboration.

1. OVERVIEW AND RELEVANCE

Rapid implementation of the Intelligent Tutoring System(ITS) and online learning has generated a large amount of students' learning data, including their writing and literacy activity data. Based on those data, recent advancements in AI and Data Science(e.g., Large Language Models and generative AI) have demonstrated the potential of EDM in tailoring instruction according to students' specific writing needs. By leveraging data collected during writing activities, educators have been able to efficiently offer targeted support and improve the learning experience. This change in literacy education encourages us to host this workshop and discuss the following themes with educators and researchers:

- Advancements in NLP for writing assessment
- Automatic essay scoring and feedback generation
- Interactive writing environments and Intelligent Tutoring System(ITS)
- Writing Dataset across disciplines
- Large Language Model(LLM) application in writing analytic or classroom practice
- Policy, privacy, and ethical concerns with writing data
- Psychometrics and measurement
- Writing behavior analysis
- K-12 writing and literacy education
- Quantitive analysis and case studies for writing education practice
- Collaborate Writing and peer review
- Writing styles and quality
- Plagiarism detection
- Data-driven supports for special education in literacy

We will announce a Call for Papers related to the above themes. In addition to regular paper presentations from the participants' submissions, we will host a tutorial session to introduce The Learning Observer, a prototype platform we are currently developing for the integration of learning data. This tool aims to set up an open, transparently-governed consortium to manage student data in the student and public interest. This open-source platform we are developing is at a prototype stage but is designed to permit the integration of diverse learning data, the use of sophisticated machine learning techniques over that data, and the presentation of real-time teacher dashboards. We are architecting it to support open science and ethical research methodology. One of our major plans is to develop a toolkit (Writing Observer) for writing education across disciplines. Specifically, we are building an open-source platform that can handle learning process data from student writing data. We can collect click-by-click online writing data on Google Docs and display real-time teacher dashboards. Therefore, we are highly interested in hearing any feedback or perspectives from the relevant researchers and educators. This tutorial session could also provide more collaboration opportunities for both us and the participants.

C. Lynch, P. Deane, P. Mitros, Z. Gao, and D. Babalola. Educational data mining in writing and literacy instruction. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 1040–1041, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12730051

We expect around 10-30 audiences for this full-day in-person workshop. Our target audience is researchers and educators with interests or experience in educational research, ITS, writing analytics, or any other related fields which highly overlap with the EDM community.

2. WORKSHOP ORGANIZERS

Collin F. Lynch is an Associate Professor in the Department of Computer Science at North Carolina State University. His primary research is focused on developing robust ITS and adaptive educational systems for Ill-Defined domains such as scientific writing, law, and software development. His current research includes work on: argument mining and natural language processing, real-time support for classroom orchestration and writing to learn tasks, advances in student modeling, the development of embodied cognitive agents for collaborative learning, and scaffolding for CS education.

Paul Deane is a principal research scientist in the Research & Development division at ETS. He is the author of Grammar in Mind and Brain, a study of the interaction of cognitive structures in syntax and semantics, and the second author of Vocabulary Assessment to Support Instruction. His current research interests include formative assessment design in the English language arts, cognitive models of writing skills, automated essay scoring, and vocabulary assessment. During his career at ETS, he has worked on a variety of natural language processing (NLP) and assessment projects, including automated item generation, tools to support verbal test development, scoring of collocation errors, reading and vocabulary assessment, and automated essay scoring.

Piotr Mitros is a Senior Research Scientist at ETS. He is also the original author of the popular Open edX learning platform and the original founder as well as the Chief Scientist for more than five years. He have spent the past few years exploring issues around why educational initiatives go south, and evidence-based practices aren't adopted and converged on issues around governance, transparency, and incentive structures. His current work focuses on how we develop educational measurements that incentivize and support rich classroom instruction supporting diverse (rather than standardized) students.

Zhikai Gao is a senior Ph.D. student at North Carolina State University. His current research focuses on understanding students' learning behaviors through traceable log data from ITS, CS education, help-seeking behavior, and LLM usage in education across disciplines.

Damilola Babalola is a second-year Computer Science Ph.D. student at North Carolina State University with a research focus on using Artificial Intelligence (Educational Data Mining and Natural Language Processing) to improve Education. His current work involves research, software development, data mining, and data visualizations aimed at assisting middle-school and high-school students in improving their essay-writing skills. The core of his research centers around the extraction and classification of student essay revisions based on their edit intention, followed by the visualization of student clusters exhibiting similar revision patterns.

3. POTENTIAL PROGRAM COMMITTEE

- Diane Litman: University of Pittsburgh
- Effat Farhana: Vanderbilt University
- Scott Crossley: Georgia State University
- Bradley Erickson: Educational Testing Service
- Danielle McNemara: Arizona State University
- Zuowei Wang: Educational Testing Service
- Rebecca Hwa: George Washington University
- Rod Roscoe: Arizona State University
- Erin Walker: University of Pittsburgh

4. WORKSHOP ORGANIZATION

We will organize this workshop as a full-day event. Half the day will be devoted to a tutorial introduction to the Writing Observer toolkit and the Learning Observer prototype platform under development at NCSU and ETS. This introduction will include a discussion of our overall architecture, an introduction to data analysis and evaluation tools, and the basic interface framework. The first half of this tutorial will be a demo of the tool we're building; we have several major components ready to present and receive feedback from the community, including a modular framework for processing learning or writing process data, a suite of NLP algorithms for extracting features from student writings, a system for rapidly developing teaching-facing dashboards, an event streaming library featuring pretty robust queuing, and more. We will work through the steps to install the system, having participants build a few dashboards and letting us know what they think, as well as what would be needed to make this useful for their own work and possible research collaborations. After this demo, depending on the remaining time, we might host a brief policy discussion. Specifically, we plan to communicate with the community about how we structure this as a consortium going forward to ensure we maintain transparency, support for diverse populations, accountability and act in the students' interest.

After lunch, the other half of the day will be devoted to paper presentations on writing and literacy instruction. We will invite submissions of full papers which describe mature work. We will also accept short papers describing in-progress work or student projects and poster/demo submissions for those presenting available data, tools, and methods. This last category is particularly targeted at researchers who have data or methods available and are seeking to identify potential collaborators.