Understanding the 'Black-Box' of Automated Analysis of Communicative Goals and Rhetorical Strategies in Academic Discourse

Elena Cotos
Iowa State University
ecotos@iastate.edu

Despite the appeal of automated writing evaluation (AWE) tools, many writing scholars and teachers have disagreed with the way such tools represent writing as a construct. This talk will address two important objections – that AWE heavily subordinates rhetorical aspects of writing, and that the models used to automatically analyze student texts are not interpretable for the stakeholders vested in the teaching and learning of writing. The purpose is to promote a discussion of how to advance research methods in order to optimize and make more transparent writing analytics for automated rhetorical feedback. AWE models will likely never be capable of truly understanding texts; however, important rhetorical traits of writing *can* be automatically detected (Cotos & Pendar, 2016). To date, AWE performance has been evaluated in purely quantitative ways that are not meaningful to the writing community. Therefore, it is important to complement quantitative measures with approaches stemming from a humanistic inquiry that would dissect the actual computational model output in order to shed light on the reasons why the 'black box' may yield unsatisfactory results.

Drawing on an ongoing project, which involves a systematic analysis of a collection of erroneous feedback produced by a genre-based AWE tool (Cotos, 2016), I will describe a hybrid – computer--driven/human-informed – approach with an exponential interpretive strand. The approach entails a linguistic investigation of the communicative goals analyzed both by AWE and the human. New heuristic taxonomies were developed to compare AWE detection and human interpretation of rhetorical intent, examine differences, and construe the nature of AWE errors. The resulting qualitative insights describe error patterns and reveal the role of linguistic features in automated detection of communicative goals. These insights help describe and interpret the reasons why error patterns in automated rhetorical analysis occur and how they may hinder computational representation of the writing construct. The findings can inform future interdisciplinary research aimed at developing augmented approaches for improving the quality of automated rhetorical feedback on student writing. In terms of immediate practical implications, the outcomes of this work can be translated to teaching and learning materials addressing possible feedback errors and providing strategies for how to use the feedback more effectively. More broadly, interpretable writing analytics can potentially power paradigmatic shifts and drive innovation at the level of research methodology, computational operationalization, interdisciplinary collaborations, and writing pedagogy – all interconnected to serve the purpose of students' writing development.

REFERENCES

Cotos, E. (2016). Computer-assisted research writing in the disciplines. In S. A. Crossley & D. S. McNamara (Eds.), *Adaptive educational technologies for literacy instruction* (pp. 225–242). Routledge: New York and London.

Cotos, E., & Pendar, N. (2016). Discourse classification into rhetorical functions for AWE feedback. *CALICO Journal, 33*(1), 92-116.