Grant R305F100007
Year of Study: 2015

**Title:** Machine learning for holistic evaluation of scientific essays.
**Authors:** Hughes, S., Hastings, P., Britt, M. A., Wallace, P., & Blaum, D.
**Citation:** Hughes, S., Hastings, P., Britt, M. A., Wallace, P., & Blaum, D. (2015). Machine learning for holistic evaluation of scientific essays. *Proceedings of Artificial Intelligence in Education,* (pp.165-175).
**Strand of work:** Tools; Design and Design-based Research on EBAIMS

**Published Abstract:**
In the US in particular, there is an increasing emphasis on the importance of science in education. To better understand a scientific topic, students need to compile information from multiple sources and determine the principal causal factors involved. We describe an approach for automatically inferring the quality and completeness of causal reasoning in essays on two separate scientific topics using a novel, two-phase machine learning approach for detecting causal relations. For each core essay concept, we initially trained a window-based tagging model to predict which individual words belonged to that concept. Using the predictions from this first set of models, we then trained a second stacked model on all the predicted word tags present in a sentence to predict inferences between essay concepts. The results indicate we could use such a system to provide explicit feedback to students to improve reasoning and essay writing skills.

**Purpose and Questions Investigated, Assessments or Tools developed**
Asking students to write essays to show learning from multiple documents is important because it is an authentic academic task requiring understanding and transforming of the content. It is a challenge, however, to score essays and to provide immediate or at least timely feedback. Automatic scoring of essays will help researchers by reducing the time required, and freeing them for deeper analyses. For teachers, it can provide important information about which students are succeeding and which are struggling.

**Research Context or Methodology**
Recent standards require students to understand explanations for phenomena, and students may need practice with feedback to learn to read and write explanations (causal models) from multiple documents. We identified four hierarchical levels of explanation quality: (1) No core content (irrelevant or vague information), (2) No causal chains (mentioned, but did not connect, elements of the causal model), (3) Causal chain with no intervening factors (elements directly linked to the outcome), (4) Chain with intervening factors (successfully included intervening elements). Humans scored high school students' essays written during two reading-to-write activities to learn about the causes of two scientific phenomena from multiple documents. Then, machine learning techniques were trained on sets of essays. For each core essay concept, we initially trained a window-based tagging model to predict which individual words belonged to that concept. Using the predictions from this first set of models, we then trained a second stacked model on all the predicted word tags present in a sentence to predict inferences between essay concepts.

**General statement of findings**
For identifying isolated ideas from source texts (concepts), the machine learning techniques are approaching the reliability of human raters (Krippendorff's alpha of 0.89 and 0.82 for two topics). For identifying quality of the explanation, we have surpassed previous attempts to infer causal relations in text (Krippendorff's alpha of 0.56 and 0.47 for two topics). The results indicate we could use such a system to provide explicit feedback to students to improve reasoning and essay writing skills.

**Implications**
This approach has the potential to be used by intelligent tutoring systems to assist students with the development of mental schema to help them refine their goals for this task, and thus aid both comprehension and production. For example, students who received the "No causal chains" feedback could be encouraged to connect the concepts they mention to the final outcome, whereas students who received the "Causal chain with no intervening factors" feedback could be instructed on how to connect these concepts via intervening concepts to the final outcome. We are currently working on identifying longer-term dependencies and coreference resolution to improve classification.

**Acknowledgments:**
The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305F100007 to University of Illinois at Chicago. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.