Using Natural Language Processing to Assess Explanation Quality in Retrieval Practice Tasks

Kathryn S. McCarthy[1] & Scott R. Hinze[2]
[1] Georgia State University, [2] Middle Georgia State University
kmccarthy12@gsu.edu; scott.hinze@mga.edu

ABSTRACT: This study explored the potential for automated assessment of students' explanations during retrieval practice. Regression analyses indicate that the linguistic features analyzed by the natural language processing tools Coh-Metrix and CRAT predicted 66% of the variance in the quality of students' retrievals. These findings indicate that both the content and connections in student retrievals are relevant to the quality of the explanation. Limitations and future work will be discussed.

Keywords: Natural language processing; reading comprehension

1 INTRODUCTION

Work in *retrieval practice* indicates that practice tests are more effective for long-term learning than restudying. Further, prompting students to *explain* what they have just read as a practice test leads to additional retention and comprehension. One such study demonstrated that students who wrote higher quality explanations during retrieval scored significantly better on a comprehension test seven days later as compared to students who merely recalled as much as they could (Hinze, Wiley, & Pellegrino, 2013). Despite the fact that open-ended practice tests are more effective than multiple-choice or fill-in-the-blank tests (Hinze & Wiley, 2011), open-ended practice tests are rarely used in classrooms due to the arduousness of providing individualized evaluation and feedback.

Thus, the current study explored if natural language processing (NLP) could be used to automate the assessment of open-ended practice tests (explanatory retrievals). Two tools were selected. The Constructed Response Assessment Tool (CRAT; Crossley et al., 2015), which calculates linguistic and semantic similarities between a source text and a constructed response was selected because it was predicted that good explanations would reflect more of the important content from the source text than poor explanations. Coh-Metrix (McNamara et al., 2014), which evaluates lexical, semantic, and cohesive features of text was selected because discourse comprehension theories assume that a more cohesive explanation is reflective of a more coherent and durable mental model (i.e., deeper comprehension).

2 METHOD

The corpus consisted of 186 retrievals collected from a study in which undergraduates (n = 62) read three science texts and then engaged in retrieval of information in each text from memory. Half of the participants were asked to *recall* and the other half were asked to *explain*, providing some variability in the quality of retrieval attempts. Two researchers scored the quality of the retrievals holistically from 1-5, consistent with how instructors typically evaluate open-ended responses (γs = .80-.89; Hinze et al., 2013 Exp. 3).

3 RESULTS

Retrievals were submitted to Coh-Metrix. Linguistic indices with non-normal distributions and those with high multicollinearity ($r > .80$) were removed. Indices that were highly correlated with quality score were retained and submitted to a stepwise regression to determine which were most predictive of the quality score. This yielded two significant linguistic indices: 1) *narrativity* (inversely related) and 2) *givenness*, a measure indicative of cohesion. The same procedure was conducted for measures in CRAT. These analyses revealed two predictors: 1) *lexical sophistication* and, 2) *semantic overlap between the source text and the retrieval*.

Finally, a hierarchical regression was conducted to determine if these linguistic indices predicted human ratings of retrieval quality. The final model, $R = .813$, $R^2 = .66$, accounted for 66% of the variance in the retrieval quality score.

Table 1: Regression analysis predicting human ratings of retrieval quality

Entry	Variables Added	R^2	ΔR^2
Entry 1	Number of Words, Text	.50	.50
Entry 2	Coh-Metrix: Narrativity, LSA Givenness	.55	.06
Entry 3	CRAT Indices: Lexical complexity (AoA), LSA Content Overlap	.66	.11

4 DISCUSSION

This exploratory study demonstrated that a combination of natural language processing tools (Coh-Metrix, CRAT) could be used to reliably predict human ratings of explanation quality in an open-ended retrieval practice. Entering indices of cohesion and content overlap significantly improved model fit, providing support for the notion that the benefits of explanatory retrieval are due not only to an increase in what is remembered, but the way that information is organized in memory.

Automating the evaluation of open-ended practice tests can make tasks like explanatory retrieval practice more amendable to classroom implementation as well as to intelligent tutoring. Given that the quality of these retrievals predicts later test performance, the ability to quickly assess what students know during practice can also serve as a form of formative feedback that instructors can use to provide remediation prior to summative tests. This study serves as an initial proof-of-concept and more will be done to improve scoring accuracy. Future work will also be conducted to replicate and generalize these findings using larger corpora on different topics. We also plan to develop and test feedback messages to help students attend to key words as well as the relations between those key words.

REFERENCES

Crossley, S., Kyle, K., Davenport, J., & McNamara, D. S. (2016). Automatic assessment of constructed response data in a chemistry tutor. In *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016)*, (pp.336-340). Raleigh, NC: International Educational Data Mining Society.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory, 19*(3), 290-304.

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language, 69*(2), 151-164.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.