EasyChair Preprint

№ 3870

Online Assessment of Students' Text Comprehension: Explorations into the Automated Scoring of Constructed Responses

Jean-François Rouet, Peter Hastings, Mônica Macedo-Rouet, Anna Potocki and M. Anne Britt

July 14, 2020

Online assessment of students' text comprehension:

Explorations into the automated scoring of constructed responses

Jean-François Rouet [1], Peter Hastings [2], Mônica Macedo-Rouet [3], Anna Potocki [1] and M. Anne Britt [4]

[1] Université de Poitiers and Centre National de la Recherche Scientifique, France

[2] DePaul University, USA

[3] Université de Paris 8, France

[4] Northern Illinois University, USA

Author note

This study was funded in part through grant ANR ANR-17-CE28-0016 (project SELEN) From Agence Nationale de la Recherche and through a Fulbright scholarship award to the first author. The opinions reported in this presentation are those of the authors. We thank Mylène Sanchiz for her participation in this study.

Abstract

Effective computerized training of reading comprehension strategies requires an ability to automatically provide accurate feedback to students as they answer open-ended comprehension questions. This study explored different approaches to machine scoring as part of a larger research and development project. Two units involving 4 open-ended questions were used for initial testing. A comparison of pattern matching and deep learning techniques suggest that the latter have more potential to provide accurate scoring of the more complex questions.

Keywords

R*eading comprehension, tutoring, feedback, machine learning*

Online assessment of students' text comprehension:

Explorations into the automated scoring of constructed responses

In this presentation we report some initial outcomes of an ongoing project whose aim was to design an online tutor to support teenagers' acquisition of purposeful reading strategies (Britt, Rouet, & Durik, 2018). More specifically, we explored different approaches to scoring students' constructed responses to questions probing their comprehension of written texts. Fluent readers construct a coherent representation of what the text says by parsing and tying together various discourse components, extracting gist structure, and drawing inferences from their prior knowledge (Kintsch, 1998). In addition, proficient reading involves an ability to use texts effectively in various contexts and for various types of purposes. Readers sometimes need to scan texts, to select relevant texts or text passages, to compare information across texts and to assess the credibility of the information (Britt et al., 2018; Leu et al., 2014). As regards school reading tasks, the use of these strategies requires an awareness of the materials and ways of knowing specific to the discipline (e.g., history, science; Goldman et al., 2016).

The SELEN project aims to develop an online system that will be able (a) to accurately assess teenagers' ability to comprehend texts in the broad sense outlined above; and (b) to manage instructional interactions with students as part of purposeful reading strategy workshops. A core challenge toward these goals is the system's ability to provide students with reliable instant feedback regarding the accuracy of their constructed responses to comprehension questions. Although there exist several approaches to the Computer-based assessment of students' written products (e.g., Hastings et al., 2019; Sabbatini et al., 2019), it is not clear which is the most effective for scoring constructed responses ranging from single words to short passages in the context of reading comprehension strategy training. In the present study, we explored the potential of deep learning algorithms vs. simpler content analysis techniques.

Method

This study is based on data collected as part of two distinct but highly similar experiments. One experiment involved college students and the other involved 8th grade middle-school students, for whom the materials were originally designed. Materials and procedures were highly similar. For purposes of simplicity we present both experiments in a single section.

Participants

The participants were first-year college psychology students (n=110) and 8th-grade students (n=550) from French schools who completed the tasks for course credit or as part of classroom assignments, respectively.

Materials

The materials were 20 reading comprehension tasks representing a range of school disciplines and other topics of interest to teenagers. Each task involved a text or a small set of texts, together with a series of 12-15 questions (see examples in Table 1). About 30% of the questions required constructed responses ranging from a single word or phrase to one or two complete sentences. The other questions were multiple-choice questions. All 20 tasks were used in the college experiment, whereas a subset of 12 tasks was used in the 8th grade experiment. The present study is based on materials used in both experiments.

Table 1

Example comprehension questions used for the machine scoring tests. The stimulus in the example unit was made of a simple chronology and a short text presenting the reign of French king Louis the Fourteenth (also known as the "Sun king").

Unit	Item	Expected response
04 - Louis XIV	How long was Louis XIV king of France?	Word or phrase (e.g., "54"; "54 years")
04 - Louis XIV	According to the text, what is a "monarchy by divine right"?	Phrase or sentence (e.g., "his power comes from God")

Procedure

The tasks were distributed into printed booklets containing 5 tasks (college, 20 units in 4 different booklets) or 3 tasks (middle school, 12 units in 4 different booklets)). Students participated in small groups (college) or in whole classes (middle schools). The completion of the experiment involved a single paper and pencil session of 60 (college) or 50 minutes (middle school). In half the sessions, students at both levels responded with the text still available to them, whereas in the other sessions they read the text first and then answered from memory. The manipulation of task context is mentioned for information only and will not be further discussed. Students' handwritten responses were typed in by the researchers for analysis.

In order to make an initial evaluation of the potential for automatically identifying the content and quality of student answers, we selected a set of answers to questions in two units, one on the history of Louis XIV (4 questions), and one on nutrition and organic food production (4 questions). For each question, a scoring rubric was developed and iteratively refined until human scorers reached a level of agreement of at least 80%. The scored answers served as input for the tagging and machine scoring of the responses.

Results and discussion

We coded the answers (approximately 150 per question) using the brat annotation tool (http://brat.nlplab.org/) for the presence of concepts which would give them full or partial credit for correctly answering the question. We performed 5-fold cross validation to evaluate the different approaches. For each fold, 80% of the coded answers were used as the comparison set, and the other 20% were used as the “test set” (as if they were new, uncoded answers). The performance on the 5 folds was then averaged to get a reliable estimate of performance on previously unseen items.

We evaluated five different methods for automatically evaluating the answers. Two methods were based on the frequency of the different codes attributed to the answers by human coders (e.g., correct, partially correct, incorrect):

Method A: Simply assign the most frequent code from the comparison set to the new answer. For example if a question received 70% of correct answers, 100% of the answers in the test set would be categorized as correct.

Method B: randomly assign a code based on the relative frequency of the different codes for that question. For example, if 50% of the answers in the comparison set for a given question were coded as correct, 25% as partially correct, and the other 25% as incorrect, this method would randomly assign codes according to that same frequency distribution.

Note that methods A and B were included mostly to acquire baseline recall and precision scores. These methods were insensitive to the contents of specific answers in the test set.

Method C was based on the identification of patterns in the answers categorized as correct or partially correct by the human scorers. Depending on the question, one, two or more patterns were identified. Any answer that included one of the predefined patterns received the corresponding code.

Methods D and E were based on a deep learning algorithm. We computed a high-dimensional vector representation of each test answer and compared that to the vector representations of the previously-scored answers from the comparison set for that question. The two deep learning variants evaluated were D) assign the classification of the pre-coded answer with the most similar vector (calculated using the cosine metric) to that of the test answer, and E) take the majority vote of the top 5 most similar pre-coded answers. The deep learning approaches used bert-as-service (https://github.com/hanxiao/bert-as-service) to compute the vectors based on the pre-trained multilingual bert embeddings released by google (https://storage.googleapis.com/bert_models/2018_11_03/multilingual_L-12_H-768_A-12.zip).

Aggregated results were calculated based on Recall, Precision, and F1 scores for both individual questions and micro-averaged over the whole set (Table 2). As expected, the deep learning and pattern matching approaches performed better than the frequency-

based approaches. The similar outcomes of the pattern matching and deep learning approaches could be attributed in part to interactions with item types. Although the small number of items tested at this stage preclude any firm conclusion, there was a tendency for simpler questions to fare better with pattern matching, whereas, more complex comprehension or evaluation questions would be best analyzed with deep learning. Table 3 illustrates this trend with two items from the "Louis XIV" unit.

Table 2

Aggregated performance indexes for the five methods included in the study.

Method	F1
A. Frequency (majority)	0.64
B. Frequency (proportional)	0.47
C. Pattern matching	0.77
D. Deep learning (closest match)	0.83
E. Deep learning (top five majority vote)	0.80

Table 3

Two example items with outcomes of the pattern matching and deep learning (closest match) approaches.

	Pattern matching Recall, Precision	Deep learning Recall, Precision
How many years was Louis XIV king of France? (item U04 Q04; type: locate; full credit : 54 years, partial credit 50 to 59 years)	Full credit .98, .95 Partial .95, .74	Full credit .96, .80 Partial .60, .38
According to the text, what is a « monarchy by divine right »? (item U04 Q06; type: integrate, full credit: right to rule is granted by God, partial credit: any response involving an act of God).	Full credit .55, .95 Partial .17, .52	Full credit .93, .73 Partial .70, .83

Discussion and perspectives

These initial results suggest that there is potential for deep learning approaches to provide accurate feedback to students' responses to open-ended questions in an online tutoring environment. Pattern matching, however, was effective for simpler questions whose correct answer may involve a number or a single word. Further tests using a larger sample will be needed to determine which particular technique provides the most accurate outcomes.

References

Britt, M.A., Rouet, J.-F., & Durik, A. (2018). *Literacy beyond text comprehension: a theory of purposeful reading*. New York: Routledge.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, MA: Cambridge University Press.

Leu, D. J., Kulikowich, J. M., Sedransk, N., Coiro, J., Liu, C., Cui, W., ... & Maykel, C. (2014, April). The ORCA Project: Designing Technology-based Assessments for Online Research, Comprehension, And Communication. *Paper presented at the American Educational Research Conference*. Philadelphia, PA.

Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M., Greenleaf, C., ... & Project READI. (2016). Disciplinary literacies and learning to read for understanding: A conceptual framework for disciplinary literacy. *Educational Psychologist, 51*(2), 219-246.

Hastings, P., Britt, M. A., Rupp, K., Kopp, K., & Hughes, S. (2019). Deep and shallow natural language understanding for identifying explanation structure. *Deep Learning: Multi-Disciplinary Approaches. Abingdon, UK: Routledge/Taylor and Francis*.

Sabatini, J., O'Reilly, T., Weeks, J., & Wang, Z. (2019). Engineering a Twenty-First Century Reading Comprehension Assessment System Utilizing Scenario-Based Assessment Techniques. *International Journal of Testing*, 1-23.