Board 65: Work in Progress: Using Natural Language Processing to Facilitate Scoring of Scenario-Based Assessments

Matthew Norris, Virginia Tech

Matthew Norris is a PhD student and Graduate Research Assistant in the Department of Engineering Education at Virginia Tech.

Mr. Hamidreza Taimoory, Virginia Polytechnic Institute and State University

Hamidreza is a Ph.D. student in Engineering Education and has a master's degree in industrial engineering at Virginia Tech (VT). He has worked in the industry as a research and development engineer. He is currently a data analyst in TLOS (Technology-Enhanced Learning And Online Strategies) at VT. His expertise is in quantitative research. His primary research interest is motivation, co-curricular activities, and professional development in engineering education.

Dr. Andrew Katz, Virginia Polytechnic Institute and State University

Andrew Katz is an assistant professor in the Department of Engineering Education at Virginia Tech. He leads the Improving Decisions in Engineering Education Agents and Systems (IDEEAS) Lab, a group that uses multi-modal data to characterize, understand, a

Dr. Jacob R Grohs, Virginia Polytechnic Institute and State University

Jacob Grohs is an Assistant Professor in Engineering Education at Virginia Tech with Affiliate Faculty status in Biomedical Engineering and Mechanics and the Learning Sciences and Technologies at Virginia Tech. He holds degrees in Engineering Mechanics (

Work In Progress: Using Natural Language Processing to Facilitate Scoring of Scenario-Based Assessments

Introduction

Evaluating socio-technical skills is a complicated and difficult task in engineering education. Scenario-based assessments have been proposed as a format providing more targeted feedback and reliable measures of student performance than existing self-report scales. Unfortunately, while these scenario-based assessments may offer more reliable measures of students' socio-technical skills, the process of scoring scenario responses remains time intensive even with trained raters and detailed scoring guides. This limits the scale at which these assessments can be deployed, and prevents adoption in the classroom. This burdensome scoring process limits scenario-based assessments' usefulness as formative assessment tools.

In this paper we are proposing a human-in-the-loop approach to assist in the classification of written responses to open-ended assessment questions. To accomplish this, we preprocess textual responses and pre-assign scores using a preexisting scoring guide. Specifically, we take responses to a scenario-based assessment and accompanying guide and utilize term extraction to categorize common terms from the response using categories from the scoring guide. Responses containing phrases that meet these scoring categories are then identified and extracted from the raw text and presented alongside that raw text to the human rater. This paper describes our process and potential benefits of its use through application in a relevant case.

The specific usage of NLP techniques presented here is simplistic and does not require detailed knowledge of computing. However, we believe that its usage to facilitate the large-scale deployment of scenario-based assessments is a useful contribution. The intentional inclusion of a human scorer remains an important aspect of the assessment-feedback process and their inclusion allows for outliers or edge-cases to be addressed more intentionally. While more advanced and comprehensive NLP techniques are widely available, we argue that more limited methods like text extraction can still provide advantages to those looking to implement these NLP in their instruction.

Background

Natural language processing refers to a range of computational techniques for the analysis of naturally developed human languages [1]. Early NLP methods utilized rule-based grammar and dictionary based frequency counts, effectively counting the number of times certain words or phrases appear in a given text. More modern methods utilize large pre-trained models [2] or transformer-based architectures [3] to address variations in semantic meaning. While advances in machine learning (ML) and neural networks (NN) have recently garnered significant attention in the business and public sphere with the release of models like ChatGPT [4] and DALL-E [5], robust applications within the field of engineering education remain are still emerging [6]. As part of the recent popularity of large language models (LLM) there have been increasing concerns about the ethical ramifications in educational and industry settings. In their analysis of the practical ethical dangers of ChatGPT Zhuo et al. [7] outline areas of concern for LLMs as a

group; the risk inherent in small models propagating with increased scale, potential biases within model training data, and the ballooning size of LLMs computational requirements. These concerns limit the number of practitioners that are willing to adopt ML, NN, or LLM tools in educational settings. Not all of these methods are appropriate or applicable to the problem at hand though; the specific NLP technique implemented must be adapted to fit the type of text being analyzed and the purpose of that analysis [8].

Existing Uses of Natural Language Processing in Assessment

With the ethical concerns and limitations of more advanced models in mind, the use of natural language processing in educational assessment is by no means a new endeavor. The use of automated essay scoring techniques have been widely developed and discussed across multiple subject areas [9]. However, their usage remains problematic and they do not see significant use in the field of engineering education.

This said, there have been many attempts at more detailed scoring mechanisms in education contexts. Smith et al. [10] discuss the development of an algorithm to automatically grade answers to open-ended reading comprehension questions for children. The algorithms developed were trained using the eBook story text, existing graded answers, and publicly accessible databases, and achieved a correct grading rate of 85%. Though this accuracy rate is high, it is insufficient for widespread use as a grading tool.

Alternatively, there exist more finely-tuned grading algorithms for assessing students' understanding. Somers et al. [11] used machine learning models to evaluate the validity of student conceptual understanding of a topic with accuracies as high as 98%. The models developed utilized four different language models and required significant knowledge of both the concepts being assessed and the technical workings of each model. More accurate and detailed models such as this require more detailed knowledge of the range of possible student answers and a prepared guide for evaluating responses.

Scenario Based Assessments

More detailed simulation-based assessments can offer a more reliable and accurate evaluation of student abilities than self-report measures [12]. However, these simulations can be time-consuming and available to only limited numbers of participants. Shorter scenario-based assessments offer authentic variable length situations that are representative of situations encountered in professional engineering.

These scenario-based assessments offer opportunities for the evaluation of behavior-based measures that require less time than detailed simulation, while providing greater reliability than self-report measures [13]. Despite these characteristics, scenario-based assessments still suffer from limitations; namely the sensitivity of individual responses to question wording [14], as well as the time still required to develop a detailed scoring guide and train human raters.

Methods

Test Case Scenario

We collected data from multiple administrations of the "Village of Abeesee" instrument, a scenario-based systems thinking assessment tool utilizing open-ended questions [15]. The scenario presents a fictional small town experiencing issues with harsh winters and hypothermic residents. Students are provided with limited information and asked to think through the process of addressing the situation. Open-ended responses to targeted questions are then analyzed by a rater and scored across seven constructs with four levels of performance (0-3). Grohs et al. [15] provide a detailed scoring guide and framework for each construct.

The majority of students achieve a score of 1 or 2; achieved through discussion of one of two broad categories related to the scenario. Through previous administrations of this instrument, raters discovered that a scoring determination of 1 or 2 was associated with specific keywords and concepts outlined in the scoring guide. This understanding and a desire to accelerate the scoring process motivated the pursuit of leveraging NLP for augmented scoring procedures.

The following is an example of the workflow for scoring the prompt "Given what you know from the scenario, please write a statement describing your perception of the problem and/or issues facing Abeesee." Scores for this prompt are determined by the inclusion of discussion of either *technical* or *contextual* aspects of the problem, with examples and categories for each aspect included in the scoring guide. Discussing either *technical* or *contextual* aspects results in a score of 1, discussing both *technical* and *contextual* aspects results in a score of 2, and a score of 3 requires students to recognize interactions between aspects.

Manual Scoring

Responses to the Village of Abeesee Scenario were collected from 162 undergraduate students in the college of engineering and college of arts and sciences. Student responses to open-ended questions were scored manually by two trained raters in accordance with Grohs et al.'s published scoring guide [15]. Scores for each response were assigned and rationales recorded. An initial sample of 20% of the responses were scored individually by each rater. These scores were then compared across raters to develop a consensus for interpreting student-generated text [16] and scoring guidelines normalized across raters. The remaining 80% of responses were split evenly between the two raters. This process required 50 human hours of work.

Facilitated Scoring

Using the RStudio and the R Shiny package we import a spreadsheet of the raw text responses. Using the scenario scoring guide and rater experience with text responses, a list of terms for scoring aspects for each systems thinking construct were developed. For example, the *contextual aspect* term list for the construct *problem identification* contains "policy, money, afford, poverty, laws, finance, history, social, politics". This list is a shortened version of the terms that indicate

students are discussing the contextual elements that may complicate the problem presented in the scenario.

Each response is then evaluated against the relevant term list for each construct, and matches highlighted for the human rater. This is completed through the use of string detection functions and regular expressions. The user can then set the score for that response, and move on to the next response in that construct. When terms from only one list are identified, the app will display a recommendation for scoring that response with a ‘1’, and recommend ‘0’ when no terms are detected. In doing this, raters are able to primarily focus on interpreting more complex responses and spend less time searching for keywords and categorizing responses.

Limitations

The process presented in this paper is by no means without limitations. The most significant limitation present is the technique of natural language processing; term extraction. Though there are methods for improving the extraction of terms [17], the applicability of the technique remains limited when compared to more modern methods utilizing language models and machine learning. This process also requires either a well-defined scoring guide or an experienced rater who has knowledge of the full range of ways in which students answer the scenario questions. This prevents this process from being useful in exploratory work or in developing an initial understanding of how students respond to questions. Although we expect this process to decrease the amount of human time and workload necessary for rating open-ended responses, the continued presence of a human rater in-the-loop restricts the speed at which answers can be rated. This was determined to be a necessary and desirable trade off, as the limitations of term-extraction as a method prohibit the interpretation of semantic meaning or novel terminology.

Potential Impact

Preliminary results indicate that this process improves the speed and reliability of scoring when compared to unassisted human scoring with the same scoring guide. Variations in the questions and complexity of scoring each construct resulted in variations in the time saved.Further testing of the process and streamlining of the Shiny app are needed to determine the extent to which this process can speed up scoring and improve accessibility.

Unlike existing auto-grader scoring systems, this process is not intended to completely offload the task of scoring from a human user. By performing term extraction and highlighting salient information our approach functionally performs a first-pass of each response, allowing a rater to focus on interpretation and outlier cases. The tradeoff taken here is that the accuracy of the model is reduced for the sake of simplicity and transparency for those looking to understand and apply NLP in a classroom. As machine learning and natural language methods continue to advance, alternative applications to augment text assessment should continue to be explored and understood.

References

[1] K. R. Chowdhary, “Natural Language Processing,” in *Fundamentals of Artificial Intelligence*, K. R. Chowdhary, Ed. New Delhi: Springer India, 2020, pp. 603–649. doi: 10.1007/978-81-322-3972-7_19.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *ArXiv181004805 Cs*, May 2019, Accessed: Dec. 15, 2021. [Online]. Available: http://arxiv.org/abs/1810.04805

[3] C. Wang, M. Li, and A. J. Smola, “Language Models with Transformers,” *ArXiv190409408 Cs*, Oct. 2019, Accessed: Dec. 15, 2021. [Online]. Available: http://arxiv.org/abs/1904.09408

[4] “ChatGPT: Optimizing Language Models for Dialogue,” *OpenAI*, Nov. 30, 2022. https://openai.com/blog/chatgpt/ (accessed Feb. 13, 2023).

[5] A. Ramesh *et al.*, “Zero-Shot Text-to-Image Generation.” arXiv, Feb. 26, 2021. Accessed: Feb. 13, 2023. [Online]. Available: http://arxiv.org/abs/2102.12092

[6] C. G. P. Berdanier, E. Baker, W. Wang, and C. McComb, “Opportunities for Natural Language Processing in Qualitative Engineering Education Research: Two Examples,” in *2018 IEEE Frontiers in Education Conference (FIE)*, San Jose, CA, USA, Oct. 2018, pp. 1–6. doi: 10.1109/FIE.2018.8658747.

[7] T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing, “Exploring AI Ethics of ChatGPT: A Diagnostic Analysis.” arXiv, Jan. 30, 2023. Accessed: Feb. 13, 2023. [Online]. Available: http://arxiv.org/abs/2301.12867

[8] J. P. Magliano and A. C. Graesser, “Computer-based assessment of student-constructed responses,” *Behav. Res. Methods*, vol. 44, no. 3, pp. 608–621, Sep. 2012, doi: 10.3758/s13428-012-0211-3.

[9] *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2003, pp. xvi, 238.

[10] G. G. Smith, R. Haworth, and S. Žitnik, “Computer Science Meets Education: Natural Language Processing for Automatic Grading of Open-Ended Questions in eBooks,” *J. Educ. Comput. Res.*, vol. 58, no. 7, pp. 1227–1255, Dec. 2020.

[11] R. Somers, S. Cunningham-Nelson, and W. Boles, “Applying natural language processing to automatically assess student conceptual understanding from textual responses,” *Australas. J. Educ. Technol.*, vol. 37, no. 5, pp. 98–115, Dec. 2021, doi: 10.14742/ajet.7121.

[12] K. Peng, R. E. Nisbett, and N. Y. C. Wong, “Validity problems comparing values across cultures and possible solutions,” *Psychol. Methods*, vol. 2, no. 4, pp. 329–344, Dec. 1997, doi: 10.1037/1082-989X.2.4.329.

[13] A. Mazzurco and S. Daniel, “Socio‐technical thinking of students and practitioners in the context of humanitarian engineering,” *J. Eng. Educ.*, vol. 109, no. 2, pp. 243–261, Apr. 2020, doi: 10.1002/jee.20307.

[14] A. F. McKenna, M. M. Hynes, A. M. Johnson, and A. R. Carberry, “The use of engineering design scenarios to assess student knowledge of global, societal, economic, and environmental contexts,” *Eur. J. Eng. Educ.*, vol. 41, no. 4, pp. 411–425, Jul. 2016, doi: 10.1080/03043797.2015.1085836.

[15] J. R. Grohs, G. R. Kirk, M. M. Soledad, and D. B. Knight, “Assessing systems thinking: A tool to measure complex reasoning through ill-structured problems,” *Think. Ski. Creat.*, vol. 28, pp. 110–130, Jun. 2018, doi: 10.1016/j.tsc.2018.03.003.

[16] J. Saldaña, *The Coding Manual for Qualitative Researchers*. SAGE, 2015.

[17] S. Aubin and T. Hamon, “Improving Term Extraction with Terminological Resources.” arXiv, Sep. 06, 2006. Accessed: Feb. 28, 2023. [Online]. Available: http://arxiv.org/abs/cs/0609019