Research Note 2022-03

Assessments in Professional Military Education: Investigating Natural Language Processing as a Substitute for Multiple-Choice Response Formats

Benjamin Nargi
Courtney Dean
Robert McCormack
Aptima, Inc.

Elizabeth R. Uhl
U.S. Army Research Institute

December 2021

United States Army Research Institute for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

Department of the Army
Deputy Chief of Staff, G1

Authorized and approved:

MICHELLE L. ZBYLUT, Ph.D.
Director

Research accomplished under contract
for the Department of the Army by:

Aptima, Inc.

Technical Review by:

Victor Ingurgio, U.S. Army Research Institute

DISPOSITION

This Research Note has been submitted to the
Defense Technical Information Center (DTIC).

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
December 2021	Final	March 2018 – May 2021

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
Assessments in Professional Military Education:- Investigating Natural Language Processing as a Substitute for Multiple-Choice Response Formats	W911NF-19-F-0020
	5b. PROGRAM ELEMENT NUMBER 633307
6. AUTHOR(S) Benjamin Nargi, Courtney Dean, Robert McCormack, and Elizabeth R. Uhl	5c. PROJECT NUMBER
	5d. TASK NUMBER
	5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Aptima, Inc. 12 Gil Street, Suite 1400 Woburn, MA 01801	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 6000 6th Street (Bldg 1464/Mail Stop 5610) Ft. Belvoir, VA 22060-5610	10. MONITOR ACRONYM ARI
	11. MONITOR REPORT NUMBER Research Note 2022-03

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

ARI Research POC: Dr. Elizabeth Uhl

14. ABSTRACT *(Maximum 200 words)*:

The current project examined the utility of using Natural Language Processing (NLP) to replace multiple-choice assessments in an existing course. A prototype tool was developed and tested. This paper provides a proof-of-concept for using NLP to replace multiple-choice assessments. Challenges to this approach are identified.

15. SUBJECT TERMS

Natural Language Processing; Assessment; Free Text Responses

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited Unclassified	13	Jennifer S. Tucker 706-366-7312

Standard Form 298

Research Note 2022-03

Assessments in Professional Military Education: Investigating Natural Language Processing as a Substitute for Multiple-Choice Response Formats

Benjamin Nargi
Courtney Dean
Robert McCormack
Aptima, Inc.

Elizabeth R. Uhl
U.S. Army Research Institute

Fort Benning Research Unit
Jennifer S. Tucker, Chief

December 2021

ASSESSMENTS IN PROFESSIONAL MILITARY EDUCATION: INVESTIGATING NATURAL LAGUAGE PROCESSING AS A SUBSTITUTE FOR MULTIPLE-CHOICE RESPONSE FORMATS

CONTENTS

Page

INTRODUCTION 1

The Current Project 1

METHOD 5

Prototype Development 5

Development of NLP for Current Effort 6

Ontology Development 8

FINDINGS 8

Challenges Identified 8

Future Research 9

CONCLUSIONS 9

REFERENCES 10

LIST OF FIGURES

Figure 1. Interface for the Prototype NLP Tool 6

Figure 2. NLP Action Flow Chart 7

ASSESSMENTS IN PROFESSIONAL MILITARY EDUCATION: INVESTIGATING NATURAL LAGUAGE PROCESSING AS A SUBSTITUTE FOR MULTIPLE-CHOICE RESPONSE FORMATS

INTRODUCTION

The U.S. Army has a variety of assessment methods in its repertoire, including the use of self-report methods and live exercises. Advancements in technology have created an opportunity for a new type of assessment which uses natural language processing (NLP) to capture and interpret realistic user responses. Brou et al. (2018) proposed the use of reactive, open-response assessments (RORAs) in response to limitations with using self-report measures and live exercises for leadership assessment in the U.S. Army. As Brou et al. noted, while live exercises may allow for more realistic and objective assessments of leadership, they are typically time consuming and costly to conduct, resulting in a limited number of evaluation opportunities for each individual Soldier. Conversely, while self-report assessments of leadership are typically low cost, in terms of time and resources, they have limitations of their own. For example, individuals may engage in socially desirable responding or may be able to identify the best response in a situational judgement test, such that an individual's response may not reflect their actual leadership skills.

RORAs present an alternative to live and self-report leadership assessments. As developed by Brou et al. (2018), RORAs present virtual scenarios that react to an individual's input. Inputs are free-text responses which are interpreted by NLP algorithms in order to determine the next step in an interactive scenario. In this way, an individual's responses are not primed by the presentation of multiple-choice options. There are challenges to developing these types of assessments, including identifying possible response options, developing appropriate branches within each scenario, and developing NLP algorithms that can properly interpret the variety of responses that may be created.

The Current Project

This project sought to extend the approach developed by Brou et al. (2018) to multiple-choice assessments in an existing virtual course. In this training course, learners are presented with computer-based personnel and operational situations where they are tasked with making leadership decisions to resolve problems. These scenarios are followed by response prompts in the form of multiple-choice options (e.g., "choose the best action from the options below"). While multiple-choice assessments are a commonly used method for extracting declarative knowledge (e.g., recall), they are not necessarily the best option for demonstrating that students comprehend or can apply that knowledge (Bloom, 1956), which are the levels of learning required for situational decision-making. That is, with multiple-choice responding students are prompted to select a doctrinal answer even when they would not have generated this answer on their own. In other words, in the absence of responses to select from, their behaviors or words may more closely approximate their decision-making processes and more authentic, unprimed responses. Further, multiple-choice options are typically set up with a right answer, a wrong answer that might closely resemble the right answer, and two distractors. The presence of the right answer can prime the student who does not recall the correct response. They may glean the

answer from the other options or use deduction to make the best educated guess. If they guess right, it is assumed they know the material and no reinforcement, remediation, or correction is made available.

A better approximation of a real-world response for students would be short answers expressing what they would say or do in the given situation. This would allow them to demonstrate their understanding of the material by applying their knowledge to solve the problem. It would allow a student to think through the problem and choose the words or actions that they deem most applicable to the situation. This has the potential for a higher degree of psychological fidelity (i.e., the student may relate more to the scenario and act as though they are in it, rather than treat it as an academic problem). The use of free text responses with NLP in place of multiple-choice questions has the potential for achieving a better, more realistic learning experience.

Natural Language Processing

NLP is a sub-field of linguistics, computer science, information engineering, and artificial intelligence that focuses on developing models and algorithms to program computers to process and analyze large amounts of language data (Garbade, 2018). An assessment system that utilizes NLP would allow students to demonstrate their responses to a situation in their own words and to receive realistic feedback on the efficacy of their responses based on the nature of the scenario branches. This would reduce the misattribution that learning occurred when the correct responding is merely the result of priming and lucky guesses. This approach would similarly be a better representation of a student's decision-making processes, as the scenarios would branch according to their own inputs.

NLP could open the door for more realistic branching and sequencing. More varied responses (than those tightly contained in four response options) could send an adaptive training environment down different paths to further explore the student's perspective. In this sense, if the student does not provide the intended answer, corrective cueing or deeper exploration into their perspective could be achieved before the student is redirected to the main branch. Further, to reduce unintended grading errors as a result of grammatical and spelling errors, a "is this what you mean?" response check could be introduced, allowing students to confirm their message prior to final submission. This system could further be used as a feedback tool to help training developers understand realistic responses to each scenario.

Testing the Utility of NLP Approaches in a Virtual Course

The goal of this effort was to test the utility of using NLP techniques for capturing realistic user responses by converting the response mechanisms for an existing virtual training course from multiple-choice to free-text inputs. Responses from the existing multiple-choice options were tied to specific branches/feedback within the unfolding scenarios in the course. A free-text input interface with underlying NLP algorithms to interpret responses was developed and inserted in place of the existing multiple-choice interface. As a first step in testing the utility of such an approach, the new free-text interface was set up to provide the same branching/feedback experiences to users.

METHOD

Prototype Development

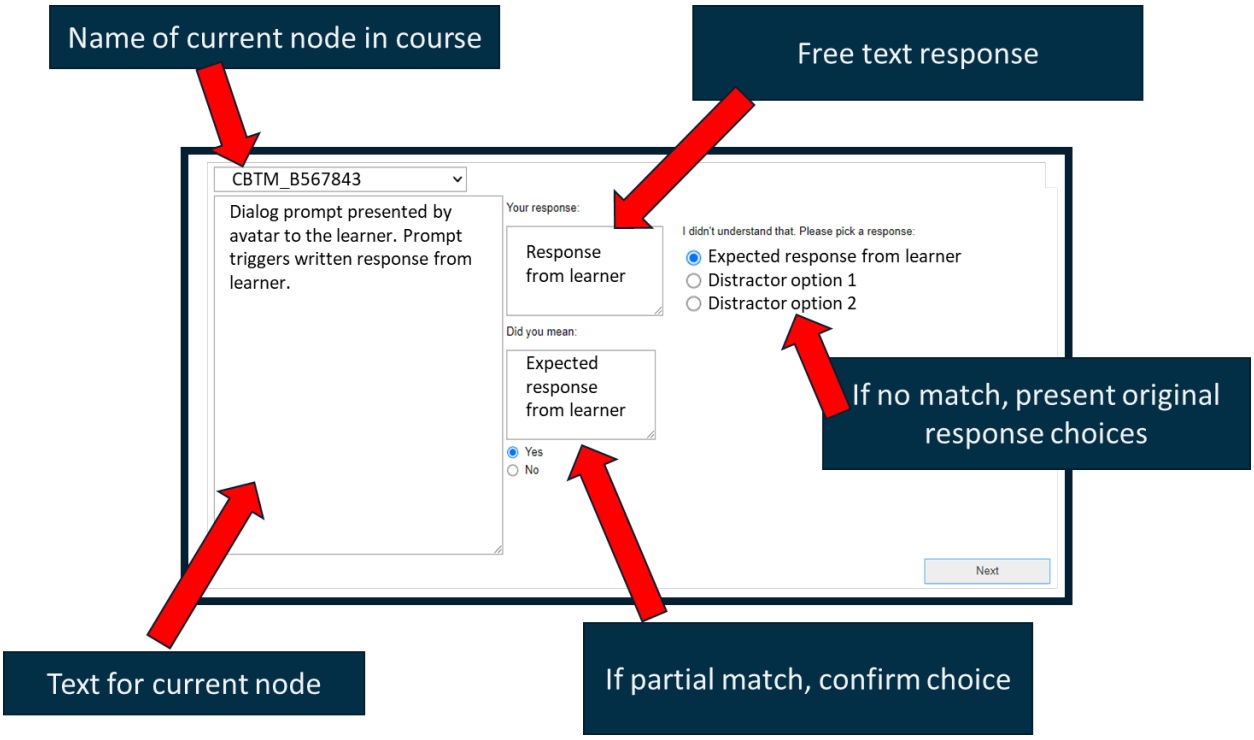
To develop a prototype of a free-text response system with NLP analysis of responses, an existing virtual course was used as the basis. That is, the content and flow of a module of an existing virtual course were exported from the course files, and the NLP analysis prototype was mapped onto the existing course structure, without making changes to the actual course. The course is structured as an online multi-media tool consisting of various scenarios in which the students respond to multiple-choice questions as though they are participants in the scenarios (e.g., squad member). Some of the responses were graded while other responses influenced the branching. That is, an individual's responses to branching questions determined the portions of the scenario (e.g., node) the individual went to next. Each multiple-choice option pointed to another node in the course flow. The course content, including the multiple-choice questions, was extracted from the existing virtual course, and the course flow was recreated in the prototype tool.

Next, a simple interface was created in the Python programming language to allow the students to see the lesson text, the questions, and the multiple-choice answers (it did not, however, include the audio or video components of the lesson). When a student was presented with a multiple-choice question, the student could select a response which then guided them down the corresponding path in the course. The multiple-choice interface was used to recreate the course flow for the NLP development.

The next step of prototype development involved altering the tool to replace the multiple-choice selection with a text box to allow the students to type in their responses. When a student typed in their response and selected the "Next" button (see Figure 1), the prototype used NLP to analyze their response. The student was then either branched to the appropriate next node based on their response or they were asked to verify their response before being routed to the next node.

Figure 1

Interface for the prototype NLP tool



Development of NLP for Current Effort

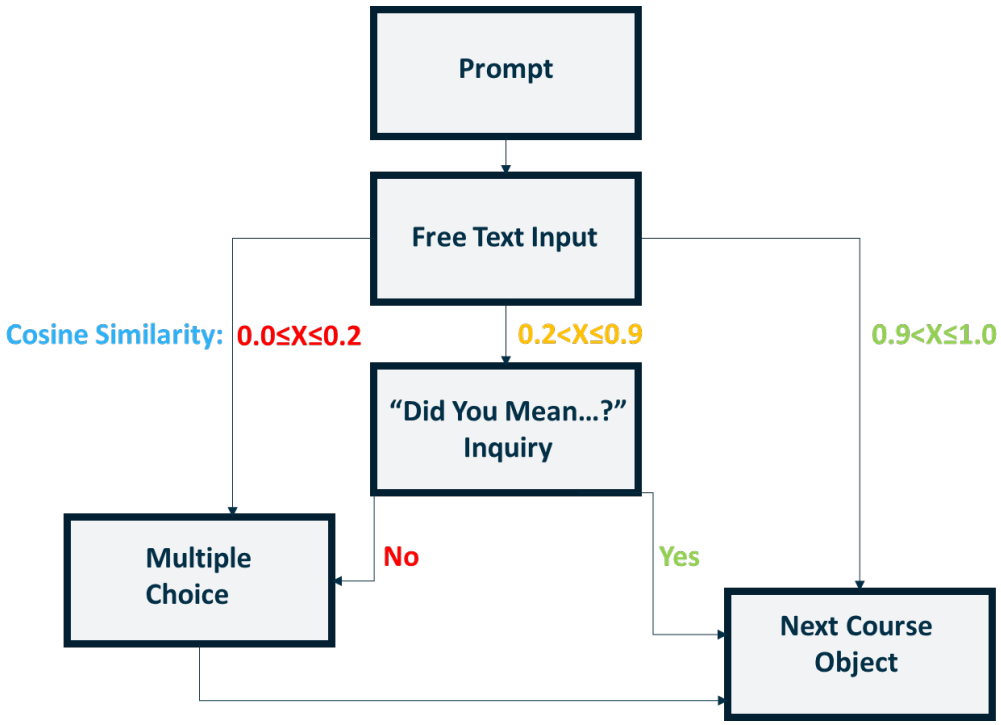
In order to process free-text responses, NLP algorithms must go through multiple steps. The first step of the NLP process was to remove "stop" words. These are words or phrases that do not convey meaning in a sentence, such as "the", "and", and "a". This allowed the NLP algorithms to focus on the words that contain the most lexical information. In addition, words were stemmed (removing "ing", "ed", etc. from the ends of the words) and depluralized (changed to singular form) to help normalize the data and remove grammatical artifacts. The result was a word vector which contained the contextual information of the students' response. That is, a list of word stems related to the response options.

The next step of the process involved matching the student's response to the closest response from the previous multiple-choice options. The multiple-choice options were similarly depluralized, stemmed, and pruned of stop words and transformed into a word vector. Matching algorithms were then used to identify which multiple-choice response was closest to the student's response. To achieve this, a cosine similarity match between the vectors was used. Cosine similarity is a comparison between two vectors which considers both the presence of words in the vector, as well as the magnitude of those words (i.e., the number of times they appear in the sentence; Singhal, 2001). In this manner, a similarity value between 0 and 1 can be assigned to the student's response and each multiple-choice response. Depending on the

similarity value, the prototype performed a different action. The flow of these actions is shown in Figure 2.

Figure 2

NLP Action Flow Chart



If the student's response (represented by *X* in Figure 2) very closely matched an existing option [cosine similarity greater than 0.9 and less than or equal to 1 (green equation)], the tool used the option as the desired response and continued the scenario with that selection. If the similarity between the student's response and the multiple-choice option was greater than 0.2 and less than or equal to 0.9 (yellow equation), the tool asked the student if that option was what they wanted to convey. If the answer was yes, the tool continued the scenario with that selection. If the answer was no, the student was prompted with the original multiple-choice options and asked to choose one. Finally, if the student's free-text response did not closely match any of the existing options [cosine similarity greater than 0.0 and less than or equal to 0.2 (red equation)], they were presented with the original multiple-choice options and asked to choose one. The interface for this tool is shown in Figure 1.

Ontology Development

The initial prototype was not robust enough to recognize substantive differences between similar responses. Thus, unless student responses closely matched the original multiple-choice options provided in the virtual course, there was a high likelihood the prototype would fail to find a sufficient match or would incorrectly match the student's response to one of the incorrect

response options. With an expanded ontology, the tool could match student responses against several alternate phrasings to achieve a higher likelihood of meeting cosine similarity. To achieve this, potential students or subject matter experts would need to analyze the current responses and develop alternative phrasings for each one; a procedure initially explored in the current effort. An initial ontology of likely potential responses was developed using personnel with behavioral science expertise. They were first asked to review the subject matter, including the scenario prompts and response options. Second, the scientists were asked to identify potential alternative ways of saying the correct and incorrect response options. The resultant lists of alternative response options were then aggregated and implemented into the prototype tool.

FINDINGS

This effort yielded a working, testable prototype to support in-depth experimentation with NLP as a means of enhancing the learner experience and more accurately assessing student learning. An existing module in a computer-based training course was modified to develop the prototype NLP tool. This allowed the researchers to test the utility of NLP for capturing realistic student responses. NLP was mapped onto an existing course structure in a prototype system that maintained the existing branching and sequencing within the course. Therefore, there was no need to develop new curriculum or fundamentally change the immersive environment.

Challenges Identified

A potential limitation of the prototype project was that the prototype had to match student responses to the original multiple-choice options, providing only a small library of alternative word choices for the NLP algorithms to compare. This was mitigated by expanding the initial ontology of potential responses for each question, with different ways of phrasing the existing multiple-choice responses. Broadening the ontology even further should yield a higher probability of a cosine match and improve the learner experience.

Second, the requirement to map free text responses to the original multiple-choice options limits the ability of students to think outside the box with their responses. In the current prototype, if the student's response was not represented by the available multiple-choice options, the student would have to then choose from the available options (see Figure 2). In order to accommodate a wide range of potential response options, pilot testing with course students or a similar population would be required to identify the most common responses so that appropriate branching options could be developed for these responses.

A related challenge is that this approach might require the development of different branching options for responses which the course leadership deemed as needing additional instructional content and feedback. In cases like this, it would be advantageous to develop responses and branching options during times when the course is being updated rather than mapping free text responses onto existing multiple-choice options. The existing multiple-choice options may not have been written to represent the most likely courses of action for each situation.

Future Research

The current project serves as a proof of concept for the use of NLP to analyze free-text responses in a virtual learning environment. Future research could evaluate the tool using the Kirkpatrick model (Kirkpatrick, 2007) by examining student reactions (Level 1) or examining the impact on learning (Level 2).

Future development could focus on the enhancement of the NLP algorithms and the response libraries in order to provide for better flow and a more realistic learning environment. Gathering responses from participant learners and SMEs would expand the response libraries and the body of possible response variations. Analysis of participant responses would enable researchers to label key words or phrases and categorize them, which would allow the development of an ontology of responses for each question. The ontology of responses could then be mapped to the existing course flow. With a more robust ontology, the NLP algorithms would be able to manage broader variations of learner responses and provide a more fluid exchange with the learners.

Further, this process may help identify new course flow requirements to address novel responses. As a broader library of possible responses from learners is developed, non-linear responses that could trigger redirection of the scenario will likely be discovered. This suggestion aligns with one of the conclusions drawn by Brou, et al. (2018) who noted that the linear nature of the RORAs used in their study reduced realism for the student as real-world conversations are seldom linear. That is, if a learner responds with statements that do not quite match the next step in the scripted dialog, the scenario could be programmed to more effectively respond to the student by redirecting the student to the scenario goal. This type of behavior would better maintain the natural flow of conversation and possibly reengage the learner. Without these additional scenario response options, the tool may miss learning opportunities with some students or students may become frustrated if the tool does not reflect their desired response.

CONCLUSION

The current research provides a proof of concept for applying NLP to free-text responses in place of multiple-choice assessments. The current paper also presents a method for testing the validity of NLP in analyzing and matching user responses to predetermined branch options. Areas for future research including building language libraries and strengthening NLP algorithms, were identified.

REFERENCES

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* David McKay Company.

Brou, R., Stallings, G., Normand, S., Stearns, I., & Ledford, B. (2018). Building automated assessments of interpersonal leadership skills. *Proceedings from Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).*

Garbade, D. (2018, October 15). *A simple introduction to Natural Language Processing*. Becoming Human. https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32

Kirkpatrick, D. (2007). *Four levels of evaluation.* ASTD Press.

Singhal, A. (2001). Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 24*(4), 35-43.