Challenges for Introducing Artificial Intelligence to Improve the Efficiency of a Next Generation Assessment Approach

SPECIAL ISSUE

Advancing Beyond Multiple Choice eAssessment

by Brian Moon, Farima Fatahi Bayat, Sneha Nair, and Andrew Slaughter

The U.S. Army sought to develop capabilities that allow for the automated or semi-automated, with greatly reduced human involvement, creation of tests and assessments. In recognizing the potential for an assessment approach that goes beyond multiple-choice, the Army chose our team to introduce and evaluate automated capabilities to author concept mapping-based assessments. This paper describes our initial approaches toward introducing efficiencies into the authoring process for concept map-based assessments. We are developing and evaluating methods to automatically generate concept maps from a knowledge domain and convert the maps into assessments for formative and summative purposes. Our initial work has sought to overcome challenges as we introduced artificial intelligence into the authoring process. In this paper, we describe our emergent approach and the challenges we have faced in seeking efficiencies in the conversion of text to concept maps.

Challenges for Introducing Artificial Intelligence to Improve the Efficiency of a Next Generation Assessment Approach

SPECIAL ISSUE

Advancing Beyond Multiple Choice eAssessment

by Brian Moon, Farima Fatahi Bayat, Sneha Nair, and Andrew Slaughter

The U.S. Army's long-term effectiveness as an organization requires that it has the ability to recruit, retain, train, and promote high-quality personnel. Key to this capability is the ability to measure cognitive abilities. For more than a century, the Army—like most learning organizations—has relied heavily on the use of multiple-choice tests as the primary means of measurement. Efficiency of deployment has been a critical enabler of this reliance, as the Army's high pace of operations demands. However, creating good assessments is neither fast nor cheap. Good, valid assessments usually require a lengthy process of close collaboration between content experts, psychometricians, and learning engineers. And even the most basic multiple-choice tests can require developers to generate huge numbers of potential items (sometimes on an ongoing basis), then review and validate them, with the process of validation sometimes being quite lengthy and expensive. These steep requirements typically entail that assessments are rarely updated or customized, tend to be deployed only in large-scale settings or high-need/high-expense contexts, are tightly controlled and often paper-based, and rarely applied to formative learning experiences or self-assessments that could potentially be very useful to individual career development.

To address these issues, the Army sought to develop capabilities that allow for the automated or semi-automated, with greatly reduced human involvement, creation of tests and assessments. The Army chose our team to introduce and evaluate automated capabilities to author concept mapping-based assessments.

Assessments based in concept maps might offer benefits for certain kinds of knowledge areas or certain applications, for example, for formative assessment. Concept maps are diagrammatic representations of knowledge in graph form that have been shown to be

effective at facilitating learning for almost 50 years [1]. They comprise organized sets of propositions, or basic statements of knowledge, that are generally organized in a semi-hierarchical shape, with the overarching concepts toward the top—the most superordinate being a single ‘super concept’—and details and more general concepts toward the bottom [2]. Thus, they show the ‘content’ and ‘structure’ of people’s knowledge and understanding, and present opportunities for assessment. Figure 1 shows an example concept map that includes the hallmark features: concepts (in boxes) linked by directed, labeled lines that form propositions, or basic units of meaning.

Figure 1. Example concept map showing hallmark features.

The U.S. Department of Education calls for concept mapping as one type of interactive computer task that is highly recommended for inclusion in every National Assessment of Educational Progress (NAEP) science assessment at the 8th and 12th grade levels [3]—primarily because concept mapping activities “tap abilities that are difficult to measure by other means.” (There are dozens of approaches—that is, tasks, rubrics—for using concept mapping as the basis for knowledge and learning assessment [4]. A meta-analysis of the research showed that methods that compare the learners’ performances on concept map-based assessments against a criterion/reference/expert/master map are the most valid and reliable method for scoring concept mapping-based assessments [5]. But despite their proven value, concept mapping-based assessments have not seen widespread adoption.

The reason is the same one the Army has noted for other assessment types: the time and labor required to manually set up an assessment, disseminate it, analyze the maps, or transcribe their content for analysis, can be considerable [6]. What Strautmene noted nearly a decade ago remains true: “There is still a need for a [concept map]–based knowledge assessment system that could perform the assessment automatically with a little intervention by the [instructor]” [7].

This paper describes our initial approaches toward introducing efficiencies into the authoring process for concept map–based assessments. We are developing and evaluating methods to automatically generate concept maps from a knowledge domain and convert the maps into assessments for formative and summative purposes. Our initial work has sought to overcome challenges as we introduced artificial intelligence into the authoring process. In this paper, we describe our emergent approach and the challenges we have faced in seeking efficiencies in the conversion of text to concept maps.

Emergent Approach

Our emergent approach merges two existing research capabilities. Sero! is an open-source concept mapping learning assessment system we developed through prior Department of Defense funding. It includes modules for manually authoring, disseminating, taking, and analyzing concept map–based assessments. Nested Information Extraction (NestIE) is an open information extraction system that extracts nested facts [8]. NestIE can detect and extract propositions within other propositions. It is based on iterative exploitation of the Iterative Memory-Based Joint OpenIE (IMoJIE) system [9].

We are merging NestIE’s information extraction capabilities with Sero!’s authoring capabilities and extending the latter as we build in algorithms to enable authoring good concept map–based assessments. The combined system first extracts propositions from text in the form of: subject–predicate→object. Next, it iteratively refines the propositions through text-based algorithms based on best practices in concept mapping. It then evaluates the emergent graph structure to offer recommendations for crafting good concept maps, also derived from best practice.

Proposition extraction. The Information Extraction (IE) field focuses on the process of extracting structured information from textual sources. Traditionally, IE systems were designed for specific domains with a fixed set of entities and relations between them. Such systems fail to generalize to various domains and have gradually been replaced with Open Information Extraction (OIE) methods. OIE systems try to automate the extraction of ontology-free textual tuples comprising relation phrases and argument phrases from within a sentence. The main purpose of these systems is to extract all facts entailed by an input text in the format of (subject; predicate; object).

Related work. State-of-the-art systems in this field are based on supervised neural approaches that can be divided into two categories. Label-based systems like RnnOIE [10] and OpenIE6 [11] treat OIE as a sequence tagging task, where each input word is labeled such that it either belongs to the subject, predicate, object, or none of the (three) main parts of a proposition. Generation-based systems such as IMoJIE [9] generate extractions sequentially using neural seq2seq models. In this work, we exploit an IMoJIE system that generates extractions using a Bidirectional Encoder Representations from Transformers (BERT)-based encoder and an iterative decoder that re-encodes previously generated extractions. This re-encoding captures dependencies between extractions and increases the overall performance.

Model architecture: Clause detector. This work relies on the IMoJIE system, and thus is a generation-based approach for information extraction. Figure 6 shows the architecture of our model. First, we feed the input text into the IMoJIE system and get the first-level propositions. An investigation of the outputs of the IMoJIE system reveals that this system can detect the subject and predicate part of a proposition with high accuracy. However, it does not perform well on extracting object segments of propositions. In fact, after extracting the subject and predicate, it labels the rest of the sentence as an object.

This drawback motivated us to process the extracted results of IMoJIE one step further. As illustrated in Figure 2, we developed a clause detection method that, given the object segment of the proposition, detects whether the object has a dependent clause within it. If there exists at least one clause in the object part, the clause’s text will be extracted and fed into the second IMoJIE system for further extraction. In this case, the outer (level 1) proposition’s object part refers to the inner (level 2) proposition(s). The system’s output is shown in Table 1.

Figure 2. Architecture of our model.

Table 1. Sample Output: Clause Detector

Input Sentence	Extraction #1	Extraction #2
Army professionals realize that they risk personal harm in performing their duty.	Ext_0: (Army professionals; realize; ext_1)	ext_1: (they; risk; personal harm in performing their duty)

Conjunction breakdown. Coordinated conjunctions such as “and” and “or” are the conjunctions that connect words, phrases, or clauses. Conjuncts in a coordinate structure exhibit replaceability [12]. This means that a sentence remains coherent and consistent, if we replace a coordination structure with any of its conjuncts. Using this property, we can break a coordination structure into its conjuncts and produce a new sentence per conjunct.

Our system is able to detect these conjunctions in different parts (effectively a subject and object part) of an extracted proposition, along with their constituent conjuncts. That is, subject and object also go through the conjunction breakdown method. Then, the proposition can be divided into multiple propositions, each with one of the conjunction’s constituents. Table 2 sheds light on this approach.

Table 2. Sample Output: Conjunction Breakdown

Input Sentence	Extraction #1	Extraction #2	Extraction #3
Enduring the physical hardship, danger, and uncertainty of combat requires an Army that is professionally committed.	Ext_0: (Enduring the physical hardship; requires; an Army that is professionally committed)	ext_1: (Enduring danger; requires; an Army that is professionally committed)	ext_2: (Enduring uncertainty of combat; requires; an Army that is professionally committed)

Sample experimental results. Table 3 shows some examples of our proposed NestIE system. Overall, we observe that the conjunction breakdown method, which detects coordination structures and produces extractions based on the number of coordination conjuncts, has high performance. Errors that arise when running this method can have several causes. One prominent cause can be the error in parsing the input sentence. Dependency parsers are powerful tools that are error-prone, and this error can propagate and affect other parts of the system.

Additionally, our system's performance in detecting and handling clauses in the object part of a proposition needs further improvement. The method that we are currently using is simple and should be developed for complicated sentence structures.

Table 3. Sample Output: NestIE preliminary results

Sentence	Extractions
War is a lethal clash of wills and an inherently human endeavor.	Ext_0: (War; is; lethal clash of wills) ext_1: (War; is; lethal clash of inherently human endeavor)
The mission, duty, and life itself require Army professionals to reject cowardice.	Ext_0: (mission; requires; Army professionals to reject cowardice) ext_1: (duty; requires; Army professionals to reject cowardice) ext_2: (life; requires; Army professionals to reject cowardice) ext_3: (mission, duty, and life; requires; ext_4) ext_4: (army professionals; to reject; cowardice)
Leaders and subordinates earn or lose trust.	Ext_0: (Leaders; earn; trust) ext_1: (Leaders; lose; trust)

Challenges. The state-of-the-art systems mentioned above are generating flat propositions and are unable to produce nested extractions. Formulating an algorithm that can extract nested propositions with relatively high accuracy has many challenges. Some of the observed difficulties in generating propositions are as follows: First, the input to our system is a single sentence, but some pronouns in the sentence have mentions that around 93 percent of them can be found within the previous three sentences. Therefore, the output extractions may have pronouns with vague mentions. Second, the correct detection of inner nests is a hard task, and there has not been much research in this area. Therefore, as we are making progress, we need to address these issues in the near future.

Refining propositions through text-based algorithms. As noted, each proposition in a concept map has three parts: subject, predicate, and object. Subject and object represent concept nodes that are linked with an edge labeled with the predicate phrase, and objects of one proposition can become subjects of others.

The main goal of this algorithm development is to get clean, sensible propositions before the concept map is generated from them. The algorithms instantiate many rules derived from best practices in concept mapping [2, 13] the end goal of which is to produce clear, concise, and coherent text for nodes and edges inside the concept map. The intention underlying the algorithms is to produce intelligent suggestions to scaffold assessment authors toward “good” concept maps that follow best practice.

The set of propositions that the concept map is generated from must have three constituents: subject, predicate, and object. If any of these parts are absent, the system needs to revise the proposition and either remove it from the propositions set or fill the absent part appropriately. If the proposition does not have a predicate, the system is currently able to generate auxiliary verbs according to the relationship between the two concepts, that is, subject and object. For example, the sentence, “Joe Biden, president of the United States, is making an announcement,” in this system has the following propositions:

(Joe Biden, is, president of the United States)
(Joe Biden, is making, an announcement)

The verb “is” in the first proposition is added by the system to the extraction.

One set of rules focuses on concepts. It includes identifying redundant concepts—singular and plural forms of the same concept, as well as concepts that are lexically and semantically similar. The system also uses pronoun resolution to make the text within a node more understandable. For example, consider the sentence, “When Army professionals

return to society, they embrace the concept of Soldiers for Life.” In this sentence, “they” refers to Army professionals. Hence the resulting extraction for the underlined part is:

(Army professionals; embrace; concept of Soldiers for Life)

Since nodes are built from text, there may be conjunctions or words starting with “wh” (such as “which”) at the beginning of a node’s text, which we want to avoid.

Concept maps should have only one super concept [2]. Thus, all concepts are checked for the edges directed towards the node (indegree). Super concepts will have 0 indegree. If more than one such concept is identified, the user will be required to choose one concept as the super concept, and new relations will be forged between this new super concept and previous super concept candidates.

Other rules involve concepts and linking phrases. First, we identify articles within each concept and remove these minor variants to reduce the size of the concept. The articles that we omit from concepts are: “a,” “an,” and “the.” The length of the concepts and linking phrases is restricted (ideally, fewer than five words) so that the text within nodes is concise. In addition, concepts and linking phrases are corrected for subject-verb agreement, which means that subject and verbal predicate of the extraction must agree with one another in number (singular or plural). Concepts may have been modified somewhere in the process and linking phrases need to be updated accordingly. If there exist concepts that are subsumed within other concepts or linking phrases, they will be extracted as stand-alone concepts and put into the concept map. Linking phrases are also checked for lexically and semantically similar phrases, in order to eliminate redundancies. Lexical similarity includes use of lemmatization and stemming [14].

Challenges. The challenges faced in this part of the system are mainly due to the lack of state-of-the-art tools available. The main issue is in detecting semantically similar phrases. Currently, popular Natural Language Processing (NLP) tools focus on identifying semantically similar words but not phrases. Phrases bring in the complexity of not only individual words, but also relationships between words. We need a clear way to mathematically compute and compare similarity scores that takes the ordering of words into consideration. In addition, similarity measures based on things like word vectors do not always adequately represent human judgments of semantic similarity. For example, failing to capture the similarity of phrases like “CEO” and “head of the organization.” However, finding singular and plural forms is made straightforward by comparing the Part-of-Speech (POS) tags from the Natural Language Toolkit library [15].

Graph-based algorithms. The main goal of this algorithm development is to convert the clean propositions produced by the above text processing into a concept map. As with the text-based algorithm, the intention underlying these algorithms is to produce intelligent suggestions to scaffold assessment authors toward “good” concept maps that follow best practice [13].

Rules for cycles, chains, crosslinks, fans and balance are computed. Balance of the map is checked only after all previous rules have been computed and the author has edited the map manually. A cycle is a path of nodes and edges (that is, concepts and linking phrases) that starts from a concept node and ends at the same node, forming an endless loop, as shown in Figure 3. Cycles are not inherently “bad;” indeed, cyclic concept maps can be quite useful [16]. However, the text-based algorithms can potentially create unintended cycles, so the goal of this rule is to point out where cycles appear and give the author the option of keeping or breaking them.

Figure 3. A concept map showing a cycle.

A chain is a path of nodes such that both concept and linking phrase nodes have no edges other than the one connecting their next neighbor; in graphical terms, chains have only one incoming and one outgoing edge, as shown in Figure 4. Chains in concept maps have been shown to express surface learning, as opposed to deep learning as expressed in network shapes [17]. The goal of this rule is to prompt authors to reduce chains in favor of more networked concept maps.

Figure 4. A concept map showing a chain.

A fan is a subgraph where a group of concept nodes (objects) have the same linking phrase and concept node (subject), as shown in Figure 5. Fans can be useful in expressing categories and families, but fans with more than three objects can suggest that intervening levels may be appropriate [18]. The goal of this rule is to reduce large fans.

Figure 5. A concept map showing a fan subgraph.

A crosslink, in concept map terms, is a link that connects two concept nodes belonging to different domains, shown in Figure 6. As Novak and Cañas [2] have noted, “Crosslinks help us see how a concept in one domain of knowledge represented on the map is related to a concept in another domain shown on the map. In the creation of new knowledge, crosslinks often represent creative leaps on the part of the knowledge producer.” Thus, crosslinks are highly valued, and the goal of this rule is to encourage the creation of crosslinks.

Figure 6. A concept map showing a crosslink.

Balance of the map means all propositions are distributed evenly, both horizontally and vertically, so that no subgraph is significantly “shorter” or “thinner” than the others in the concept map, as suggested in Figure 7. Highly unbalanced maps might suggest that sections of the map require additional attention or perhaps should become new maps [18]. Thus, the goal of this rule is to focus attention on imbalance.

Figure 7. A concept map showing imbalance.

We have faced challenges introducing these rules that are due to the variable nature of concept maps, which is desired, and the interactions amongst the rules. For example, in the balance rule, the height of a subgraph needs to be computed to compare across other subgraphs. However, the subgraph may contain cycles, resulting in endless loops. To check the balance of the map, individual branches stemming from the super concept need to be considered. But sometimes, cycles link such branches, which creates a problem as the balance rule will incorrectly view such subgraphs as a single branch instead of multiple individual branches. Crosslinks also create challenges for computing balance, as the algorithm also takes into consideration links to other subgraphs. Crosslinks are also hard to identify as such. They are not the same as the cross edges in depth-first traversal of a tree. They are also not strong bridges, as their removal does not result in strongly connected components, which have paths between every pair of nodes so that every node is reachable from every other node [19].

We have implemented algorithms to process the text- and graph-based components. Other than the super concept rule, all of the rules in the text and graph services are presented as optional to the assessment author.

Additional and Future Work

In addition to the text-to-map efficiencies, we are also implementing a set of algorithms that offer recommendations for which types of concept map-based assessments are appropriate for given maps, how difficult maps might be for takers, and which assessment items are appropriate for the skeleton map type. Future work will include user testing of the entire approach and comparing it against other authoring approaches, including other information extraction approaches and manual authoring, with the ultimate goal of introducing significant efficiencies into assessment authoring.

Acknowledgement

This material is based upon work supported by the Army Research Institute for the Behavioral and Social Sciences and the Army Contracting Command under Contract No. W911NF-20-C-0028.

References

- [1] Novak, J. D. *Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations*. Routledge, New York, 2010.
- [2] Novak, J. D. and Cañas, A. J. The theory underlying concept maps and how to construct them. Florida Institute for Human and Machine Cognition. 2006.
- [3] National Assessment Governing Board. [*Science Framework for the 2015 National Assessment of Educational Progress*](#). 2014.
- [4] Ruiz-Primo, M. Examining concept maps as an assessment tool. In Cañas, A., Novak, J., et al. (Eds.), *Concept Maps: Theory, Methodology, Technology. Proceedings of the First International Conference on Concept Mapping*. Universidad Pública de Navarra, 2004, 555–562.
- [5] Himangshu S. and Cassata-Widera, A. Beyond individual classrooms: How valid are concept maps for large scale assessment? In *Proceedings of the Fourth International Conference on Concept Mapping*. Universidad de Chile, 2010.
- [6] Cañas, A. et al. CmapAnalysis: An extensible concept map analysis tool. In *Proceedings of the Fourth International Conference on Concept Mapping*. Universidad de Chile, 2010.
- [7] Strautmane, M. Cmap-based knowledge assessment tasks and their scoring criteria: An overview. In *Proceedings of the Fifth International Conference on Concept Mapping*. Institute for Human and Machine Cognition, University of Malta, 2012.

- [8] Bhutani, N. et al. Nested propositions in open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016, 55–64.
- [9] Kolluru, K. et al. IMoJIE: Iterative memory-based joint open information extraction. arXiv preprint arXiv:2005.08178. 2020.
- [10] Stanovsky, G. et al. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers). Association for Computational Linguistics, 2018, 885–895.
- [11] Stanovsky, G. and Dagan, I. Open IE as an intermediate structure for semantic tasks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Volume 2 (Short Papers). Association for Computational Linguistics, 2015, 303–308.
- [12] Saha, S. Open information extraction from conjunctive sentences. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 2018, 2288–2299.
- [13] Moon, B. et al. (Eds.). *Applied Concept Mapping: Capturing, Analyzing, and Organizing Knowledge*. CRC Press, 2011.
- [14] Stanford NLP. [Stemming and lemmatization](#). 2008.
- [15] [Natural Language Toolkit](#). 2021.
- [16] Safayeni, F. et al. A theoretical note on concepts and the need for cyclic concept maps. *Journal of Research in Science Teaching* 42, 7 (2005), 741–766.
- [17] Hay, D.B. and Kinchin, I.M. Using concept maps to reveal conceptual typologies. *Education + Training* 48, 2/3 (2006), 127–142.
- [18] Moon, B. M., et al. Skills in applied concept mapping. In Moon, B., Hoffman, R. R., Novak, J., and Cañas, A (Eds.), *Applied Concept Mapping: Capturing, Analyzing, and Organizing Knowledge* CRC Press, 2011, 23–46.
- [19] Maurer, P. M. Generating strongly connected random graphs. In *Proceedings of the International Conference on Modeling, Simulation and Visualization Methods (MSV)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing, 2017, 3–6.

About the Authors

Brian M. Moon is the chief technology officer for Perigean Technologies, president of Sero! Learning Assessments, and cofounder of BMC, along with Jeff Ross and Martyn Roads. His research interests include expert decision-making in naturalistic environments, and the assessment of mental models, particularly using concept-mapping techniques.

Farima Fatahi Bayat is a first-year graduate student of the Computer Science and Engineering department at the University of Michigan. She is a research assistant who is advised by Prof. H. V. Jagadish and is working in the areas of information extraction and data mining. Bayat holds a Bachelor of Science from the University of Tehran, Iran, where she worked with Dr. Mehdi Modarressi in the Parallel and Network-based Processing Research Group.

Sneha Nair is a software engineer at Perigean Technologies. She is leading development of automated methods for authoring concept mapping-based assessments. Nair holds a Master of Science in Computer Science from University of Texas at Dallas. Her main interests are in the field of Natural Language Processing and software architecture.

Andrew Slaughter is a scientist in the Predictive Analytics Research Unit at the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI). His primary research interests include social network analysis, psychometrics, and the study of individual differences. He has a doctorate in organizational psychology from Texas A&M University.

DOI: 10.1145/3470745