A Novel Approach to Automated Assessment Generation Using Semantic Extraction

Terry Patten[1], Ph.D., Rachel Amey[2], Ph.D., Joanne Barnieu[3], Clarence Dillon[3], Jennifer Harvey[3], Ph.D., Sean Shiverick[3], Ph.D., Michael Smith[3], Steve Hookway[1]

1: Charles River Analytics, Cambridge, MA (tpatten@cra.com, shookway@cra.com)

2: U.S. Army Research Institute for the Behavioral and Social Sciences, Fort Belvoir, VA (rachel.c.amey.civ@army.mil)

3: ICF, Reston, VA (joanne.barnieu@icf.com, clarence.dillon@icf.com, jennifer.harvey@icf.com, sean.shiverick@icf.com, mike.smith@icf.com)

ABSTRACT

Rapid and reliable individual-level assessments are critical to developing effective and capable Army Soldiers and meeting the needs of modern warfighters. Changes in required skillsets or equipment require more frequent updates to training and the creation of new training courses, which, in turn, leads to creation of new assessments to measure knowledge, skills, and abilities (KSAs). Traditional methods of assessment development involving manual item construction are labor-intensive, time-consuming and often costly. The Army is investigating methods to scan available training text (e.g., field manuals, lesson plans) to automatically generate assessments, reducing the need for human involvement. While syntactic and neural network methods for automated assessment are currently prevalent, the team is applying novel semantic information-extraction technology to address the limitations of these approaches. Using the frame-based approach pioneered by Berkeley, it is possible to generate a variety of grammatically and semantically sound stems and response options unconstrained by the original wording in the instructional material. Additionally, this approach does not require large amounts of language-model training data, is capable of assessing higher levels of reasoning, and allows model parameters to be adjusted directly. The research team has proven the feasibility of this automated approach and developed a system prototype which generates items from test-bed training source material concerning a type of Army equipment.

ABOUT THE AUTHORS

Terry Patten, Ph.D. is a Principal Scientist at Charles River Analytics, focusing on the analysis and generation of natural language text and the computational modeling of linguistic theory. His primary research interest is computational linguistics (natural language processing) including information extraction, semantic technology, natural language generation, sentiment analysis, computational modeling of the social aspects of communication, text mining, and machine learning of linguistic knowledge He received his Ph.D. in Artificial Intelligence from the University of Edinburgh.

Rachel Amey, Ph.D. is a Research Psychologist in the Predictive Analytics and Modeling Research Unit at the U.S. Army Research Institute (ARI). Her ARI research focuses on how to apply machine learning techniques to extract key information from natural language data to understand hidden themes. Her primary research interests are human cognition, learning and memory, machine learning, and network analyses. She received her Ph.D. in Psychological and Brain Sciences from the University of Delaware and completed a post-doc in computer science at Drexel University.

Joanne Barnieu, M.S. is a Senior Lead Learning Scientist at ICF, focusing on the design and execution of military training effectiveness research studies. Ms. Barnieu has over 30 years of experience in instructional systems design and specializes in the use of innovative training and assessment strategies. Ms. Barnieu is bilingual English/French and holds a B.A. in French Education from Franklin and Marshall College and an M.S. in Organizational Development from Saint Joseph's University.

Clarence Dillon, M.S.S. is a Senior Data Analyst at ICF with a background in computational social science and data analysis with a career interest in analysis of complex, data-intensive problems that require computational solutions. He holds a Master of Social Science from the University of Tampere, Finland with a thesis on formal modeling of foreign policy decision-making using deep textual analysis.

Jennifer Harvey, Ph.D. is a Director of Human Capital at ICF with 20 years of personnel research and consulting experience with private, public, and military organizations. Dr. Harvey's skills include project management, test development, qualitative data collection and analysis, descriptive and inferential statistical analysis, experimental and survey research, and technical writing. She has developed a variety of content-valid personnel assessments, including text-based and computer-animated situational judgment tests. Dr. Harvey received her Ph.D. in Industrial/Organizational Psychology from the University of Akron.

Sean Shiverick, Ph.D. is a Research Psychologist at ICF with 15 years of experience in analytics, research designs, and behavioral science. Dr. Shiverick's skills include quantitative and qualitative research, statistical analysis, and data science methodologies. He has supported client research on assessment development and validation, natural language processing (NLP), and analysis of biometric data for stress detection and instructional innovation. Dr. Shiverick received his Ph.D. in Psychology from the University of Wisconsin and M.S. in Data Science from Indiana University.

Michael Smith, M.P.P. is a Senior Director of Data Analytics at ICF with more than 15 years of experience in strategic planning, analytics, risk assessment, project management, and innovation, including 14 years of experience supporting a wide variety of planning efforts for the Department of Defense. Mr. Smith currently advises several clients on how to adapt emerging data science practices to improve organizational performance. Mr. Smith holds a Master of Public Policy (M.P.P.) from Georgetown University.

Steve Hookway, M.S. is a Senior Software Engineer at Charles River Analytics, where he leads the research of semantic web technologies. Prior to joining Charles River Analytics, Mr. Hookway worked as a Software Engineer for Lockheed Martin's Advanced Technology Labs. He has a B.S. and an M.S. in Computer Science from Lehigh University.

A Novel Approach to Automated Assessment Generation Using Semantic Extraction

Terry Patten[1], Ph.D., Rachel Amey[2], Ph.D., Joanne Barnieu[3], Clarence Dillon[3], Jennifer Harvey[3], Ph.D., Sean Shiverick[3], Ph.D., Michael Smith[3], Steve Hookway[1]

1: Charles River Analytics, Cambridge, MA (tpatten@cra.com, shookway@cra.com)

2: U.S. Army Research Institute for the Behavioral and Social Sciences, Fort Belvoir, VA (rachel.c.amey.civ@army.mil)

3: ICF, Reston, VA (joanne.barnieu@icf.com, clarence.dillon@icf.com, jennifer.harvey@icf.com, sean.shiverick@icf.com, mike.smith@icf.com)

INTRODUCTION

Rapid and reliable individual-level assessments are critical to developing effective and capable Army Soldiers and meeting the needs of modern warfighters. Changes in required skillsets and new or changing equipment require more frequent updates to training and the creation of new training courses, which, in turn, leads to creation of new assessments to measure knowledge, skills, and abilities (KSAs). Traditional methods of assessment development involving manual item construction are labor-intensive and time-consuming and can be costly to implement on a recurring basis. For example, as compared with essays or other response-constructed tasks, multiple choice (MC) items are commonly used in educational exams or training assessments due to the efficiency in administering the exam, the objectivity in scoring, and the time spent by students recording answers. While MC items may be efficient to administer and score, they are time-consuming for the instructor to develop, since the incorrect response options require a content specialist to create plausible yet incorrect answers. When an instructor develops 100 four-option MC items, he or she creates 100 stems and 100 keys but is required to create 300 incorrect responses, known as distractor responses (Gierl, et al., 2017). The Army is interested in ways to automate the assessment development process to reduce this time and effort involved in manual item creation. Development of an automated assessment tool which scans training source content, extracts important semantic relationships, and generates a large quantity of items (stems, keys, and distractors) based on those relationships would solve time and labor-related challenges associated with assessment development.

AUTOMATED APPROACH

Three current techniques for automated item generation (AIG) include syntactic approaches, neural network approaches, and semantic approaches. The syntactic approach to AIG involves formal or structural transformation of a sentence that alters the grammatical structure of a sentence to produce an assessment item, such as by changing a declarative sentence into an interrogative sentence (Le et al., 2014). Additionally, the use of neural network techniques is rapidly increasing (Götz, et al. 2021). These techniques, used for machine translation, can transform statements into questions based on very large training sets of statement-question pairs (Subramanian et al., 2017; Zhou et al., 2017). By contrast, semantic approaches to AIG extract the meaning or semantics from source sentences and generate items from deeper conceptual-level relationships, rather than the syntax of the input (Kurdi, Parsia, & Sattler, 2017). A comprehensive review of AIG literature focused on studies involving item generation for educational purposes (since 2015), indicated that most item generation approaches (65%) were based on semantic information and a smaller proportion (11%) based on syntactic information (Kurdi, et al., 2019). The characteristics of each approach have advantages and disadvantages for generating the stem, key, and distractors.

Syntactic Approach

Syntactic-based approaches transform syntactic features of the input (e.g., parts of speech) to guide question generation, but do not require semantic understanding of the input (Le, et al., 2014). A simple case involves replacing the salient information in the sentence with a blank (i.e., fill-in-blank questions) or response option for a MC item. A more advanced case involves generating "Wh-" questions (e.g., who, what, why, where, when) by transforming the source sentence with basic substitution rules (Ch & Saha, 2018). For example, rules could be written to transform instructional sentences such as "Fleming discovered penicillin" into item stems such as "Who invented penicillin?"

or "What did Fleming discover?" An advantage of the syntactic approach is that no additional information is required to generate the question, as all the material is present in the source material (i.e., the training content). A disadvantage to the syntactic approach is that the complexity of the source material or a reference to information external to the source can produce questions that are incorrect, confusing, or incoherent. Another disadvantage is the wording of the assessment items is constrained by the wording of the original sentence (e.g., if the original sentence is passive voice, the assessment item stem will be passive voice). Furthermore, finding reasonable distractors (i.e., plausible incorrect answers) is very time-consuming. Across different approaches to AIG, most existing methods for distractor selection involve various similarity measures (Liang, et al. 2018).

Neural Network Techniques

Neural network models have been adopted for AIG using techniques such as word embeddings or vector embeddings (Le & Mikolov, 2014; Mikolov et al. 2013). An advantage of the neural network approach is that, given sufficient training data, the transformation rules are learned automatically without requiring additional resources. Two disadvantages of neural networks are (1) they require a massive amount of training data and (2) there is no way to refine a neural network other than trying a different learning algorithm or obtaining better training data—the former is time-consuming for developers, and the latter can be challenging to obtain. Previous research has used classification models or machine learning ranking models for distractor selection (Liang, et al. 2018); however, these methods require more data and comparing the performance of several computational models takes more time.

Semantic Approach

In past research, semantic information has been represented using ontologies, which are knowledge structures that define concepts, entities, and their relationships (Papasalouros, et al. 2008; Vinu & Kumar, 2015). For example, a radio is an electronic communication device; a radio has an antenna and a power source; a radio has a transmitter and a receiver, etc. In an ontology, terms and concepts are organized in a hierarchy of classes that provides a framework for generating item stems and response options based on the relations between classes or properties (Cubric & Tosic, 2010). The primary limitation of the ontology-based approach to AIG is that all the knowledge assessed must be represented in the ontology, which can require considerable time and effort to build when an existing ontology is not available.

The semantics of sentences can be also expressed using *thematic roles* such as an Agent who performs an action or the Location where the action is performed. Determining the semantic roles in a sentence reveals the meaning of the sentence—who did what to whom, where, and when (Flor & Riordan, 2018). Importantly, these thematic roles do not change across grammatical variations such as active versus passive voice. Once the thematic roles for any specific semantic relationship have been determined, item stems can be generated automatically and independently of the original sentence. Following the example above, once the thematic roles have been extracted from any of the variations, many different item stems can be generated (See Table 1).

Table 1. Thematic Roles and Item Stems

Source Sentences for the Thematic Roles: Agent=John, Beneficiary=Mary, Theme=the book	**Item Stems that can be Generated from any of the Source Sentences**
John gave the book to Mary. Mary was given the book by John. The book was given by John to Mary. John gave Mary the book. It was John that gave the book to Mary. What John gave to Mary was the book.	What did John give to Mary? Who gave the book to Mary? To whom did John give the book? John gave the book to _____. True/False: John gave Mary the book. The book was given to Mary by whom? John gave Mary the _____.

Assessment developers can generate parallel forms of an assessment using the semantic approach because it can produce a large number of alternate items. These examples (Table 1) assess basic recall of a specific semantic relationship. By generating "how" or "why" questions from the same semantics, the technology can generate items that assess higher reasoning or cognitive levels, beyond simple recall, based on inferences such as (1) Mary has the book. How did she get it? (e.g., understanding), and (2) If Jane wants the book, what will she need to do? (e.g., apply).

Higher levels of reasoning can also be assessed by developing items based on different semantic relationships, such as troubleshooting, that ask a learner to reason about the possible remedies for a stated problem (Leo, et al. 2019).

Unlike the syntactic approach, generating items from semantics is not constrained by the original wording and provides the flexibility to modify the type or format, readability, difficulty, and cognitive or reasoning level of the generated items. Items can be generated from semantic information that was not derived directly from instructional text. For example, the semantic information could come from a table in a manual, from a database, or by aggregating information extracted from several different chapters in a manual. The semantic approach does not require large training datasets, as required by neural networks or other statistical language models. The potential disadvantage of the semantic approach is that building the lexicon and grammar require significant effort by a computational linguist. However, the economics of the semantic approach are determined by scale: the initial time and investment involved in building the grammars used to generate item stems can be recouped by adapting the grammars to new content areas with relatively low human effort. Therefore, the semantic approach can be economical if the costs are spread over many courses, as is the case with the Army and other large-scale educators.

Based on the analysis of existing approaches to automated assessment, the research team studied the feasibility of building an automated assessment prototype using semantic technologies and then went on to build the prototype, which would generate complex and varied items at a large scale. These technologies are described in the next section.

AUTOMATED ASSESSMENT TECHNOLOGIES

Specific technological requirements for the semantic approach include computational grammars of English that relate English sentence constructions to semantics, processing technology that can apply these grammars to extract the semantics from the text, and processing technology that can generate assessment items from semantics. This section describes the off-the-shelf technologies adopted for the current research—FrameNet, Systemic Functional Grammars (SFGs), and the SFG Toolkit. The SFG Toolkit was previously developed for the Defense Advanced Research Projects Agency (DARPA). These technologies work together to constitute the semantic approach. FrameNet and SFGs represent the linguistic theory, and the SFG Toolkit processes the instructional text using the grammars. The SFG Toolkit finds and extracts instances of specific semantic frames by matching the frames described in the grammar against the input text.

FrameNet

Since Army instructional material contains many semantic concepts and relationships, a potential disadvantage to the semantic approach for Army use is that defining all the semantic relationships may be too labor-intensive. This issue can be mitigated, however, by using an existing repository of semantic relationships. Many relevant semantic roles and relationships are provided by the FrameNet project at the University of California, Berkeley (Ruppenhofer et al., 2006; https://framenet.icsi.berkeley.edu). FrameNet defines a *frame* for many semantic relationships, which includes a frame element for each participant in the relationship corresponding to a thematic role. Extracting a frame from a sentence involves automatically determining which sentence constituents correspond to which frame elements. Once an instance of a frame is extracted from the instructional text, it is straightforward to generate various item stems that ask about its frame elements. Importantly, instructional material typically focuses on a relatively small set of semantic relationships: For example, equipment manuals focus on the names of the components of the equipment, the purpose of the components, how to assemble, configure, and operate the components or system as a whole, and how to troubleshoot components. These are the semantic relationships that must be learned to use the equipment effectively.

The purpose of equipment components, for example, is addressed by FrameNet's *Tool_Purpose* frame (See Figure 1), which describes the semantic relationship between two core frame elements (FEs in Figure 1), the *Tool* and its *Purpose*. Once an instance of this frame is extracted from the instructional material, items can be generated to ask about any of the frame elements. Since the semantics of the relationship are known in advance, the wording of the stem does not have to follow the wording of the original sentence. Suppose a field manual states that, "Round nosed pliers are used to make screw loops." Here, the *Tool* is "round nosed pliers," and the *Purpose* is "to make screw loops." Once this frame is extracted, a variety of item stems can be generated, including:

- What is the purpose of round nosed pliers?
- What are round nosed pliers used for?
- What is used to make screw loops?

- Screw loops are made using _____.

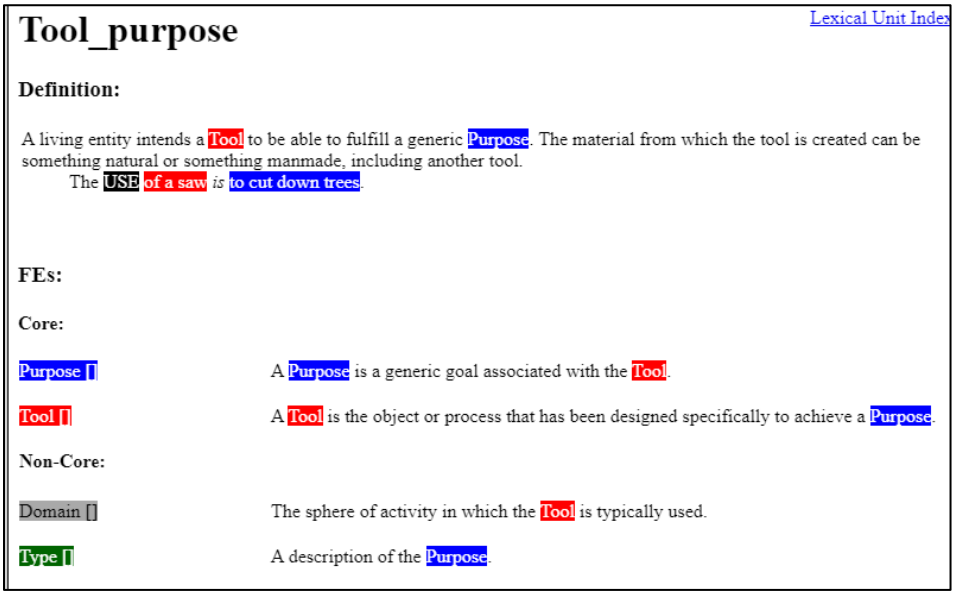

Figure 1. *Tool_Purpose* Frame from FrameNet
(https://framenet.icsi.berkeley.edu)

Frames and Distractors

One of the results of the research is that this frame-extraction approach to item generation provides a unique and surprisingly effective technique for discovering distractors. The frame elements of each frame have specific semantic properties. The *Tools* are components of the equipment, and the *Purposes* are specific goals they achieve. This means that *Tools* from extracted *Tool_purpose* frames tend to be good distractors for each other, and *Purposes* from *Tool_purpose* frames tend to be good distractors for each other. As a result, the frame-extraction approach to stem generation has the significant advantage of producing good distractors (for other stems) as a side effect, alleviating the time and labor-related challenges associated with the syntactic or neural network approaches, which require separate mechanisms to be constructed to find distractors.

Systemic Functional Grammars (SFGs)

Extracting semantic frames from field or technical manuals does not require full understanding of the complete sentences from a manual, only the ability to recognize specific key semantic relationships relevant to the type of instruction. Therefore, the type of grammar best suited for information extraction is one that is capable of both partial parsing and semantic analysis. SFGs (Halliday, 2003; Halliday, 2007; Halliday & Matthiessen, 2014) are a type of grammar that meets these criteria. SFGs are heavily oriented toward semantics rather than syntax. They also are a constraint grammar, so they are easily used for partial parsing by ignoring some language constraints. SFGs have been used in seminal computational systems for both language analysis (e.g., Winograd, 1971) and language generation (Mann & Matthiessen, 1983). It is the ability of SFGs to assign thematic roles that makes them ideal for supporting the semantic approach to automating the generation of assessment items. SFGs provide an economical and easy-to-understand method for relating field or technical manual text to the underlying frame semantics needed for assessment item generation.

SFG Toolkit

Given that SFGs provide the type of grammar necessary to support the semantic approach to item generation, technology that processes SFGs is required. The research team employed the Government-funded SFG Toolkit, developed for DARPA projects, which supports both text analysis and text generation using SFGs, and the integration of SFGs with FrameNet semantics. The SFG Toolkit has two primary components: The SFG Builder, which is a

graphical editor used to create and edit the grammars; and the SFG Engine, which uses the grammars to computationally analyze and generate text. Once an SFG is created using the SFG Builder, it can automatically extract the frame semantics from sentences appearing in the instructional material. The accuracy of the automatic extraction depends on the complexity and consistency of the language in the instructional material.

TEST-BED CONTENT

For the automated assessment research, the team decided to leverage instructional content from a previous but unrelated applied research project involving a standard combat radio (Spain, et al., 2013; Long, et al., 2015). This previous research involved the development of an immersive training course for leaders at the Army Signal School and included a virtual, interactive radio. In addition to research questions surrounding learner engagement and training effectiveness, the team studied the impact of integrated assessments in the interactive training course, including a computer adaptive test and checks on learning. Due to this previous research, the automated assessments research team had access to the radio operator manual as well as a bank of assessment items generated by humans and validated by radio subject matter experts. Some of these items were evaluated using student assessment results and were shown to be predictive of actual hands-on performance with the radio. These resources allowed the team to use the operator manual as the source content for investigation of the semantic technologies. Since the radio is a piece of equipment, it made sense to begin with the Purpose frame (as seen in Figure 1) to test the performance of the semantic technologies as part of the prototype development. The research team engaged in frequent item review sessions to determine if the automatically generated items (stems, keys, and distractors) were viable or if they contained flaws that needed to be addressed.

After conducting initial feasibility studies, the team embarked on the automated assessment research to develop and refine the prototype and item output. After several cycles of iterative prototype development using the combat radio operator manual, the research team tested the generalizability of the technology by applying it to a different type of instructional material for the combat radio (i.e., lesson plans). One of the challenges included bulleted lists, since the lesson plans were created by an instructor who summarized key points from the operator manual. The technology was still able to extract semantic frames from the bulleted lists. Further, during the item review sessions, the team determined the lesson plan source material led to more viable items. However, the instructor involvement in the content curation could have been a factor in this result. The instructor determined which content was most salient to pull from the operator manual. Further, the instructor simplified the language in many cases when summarizing the information.

The team then moved on to a different subject matter (i.e., an electrical systems field manual) to test the generalizability of the technology to different subject matter. The team was interested in determining what level of effort would be needed to produce items from a new manual, while levering the *Tool_purpose* grammar and lexicon developed for the combat radio operator manual. The next section outlines the research methods and provides results.

RESEARCH METHODS

To demonstrate the feasibility and utility of the frame-semantic approach to automated item generation, the team used the technology to extract *Tool_purpose* frames from a technical manual relevant to electrical systems and then generated stems and distractors for assessment items based on those frames. The team evaluated the results through a series of item review sessions. One of the feasibility issues addressed is the amount of grammar and lexicon building that is required for the approach. The team assumed that, in practice, the grammar and lexicon merely need to be adapted from previous lexicons and grammars, rather than starting from scratch. Therefore, the team made only minor adjustments to the radio grammar and lexicon prior to applying them to the electrical systems manual.

Process

The PDF file for the electrical systems manual was converted to a plain text file. A small amount of effort was then put into cleaning up PDF artifacts such as removing the table of contents, page headers, etc. Automatic entity recognition software was then run to find multi-word entity names. At this point, no manual modifications had been made to the system's grammar or lexicon. The automated item generation system was then run on the electrical systems manual to produce *Tool_purpose* item stems, and candidate distractors, which were the corresponding frame elements of other extracted *Tool_purpose* frames. The system was run initially on the electrical systems manual with

no modifications to the grammar or lexicon. Several days were then spent making minor adjustments to the lexicon and grammar and the system was run again. The goal of this exercise was to determine the amount of work that is necessary when preparing the technology for a particular course.

The system outputs a set of multiple-choice items, each consisting of a stem, the correct response, and a list of candidate distractors. Typically, multiple choice assessment items include three distractors. The prototype system outputs a longer list consisting of the corresponding frame elements from other extracted frames. Some of the distractors may be inappropriate for subtle technical reasons that are beyond the technical understanding of the system, but which human experts can easily see based on their deeper knowledge. Our hypothesis is that the instructor or assessment designer can pick the best distractors from the list quickly, so the technology still provides a significant time savings.

Results

The initial run produced 121 items, of which only 13 were usable without refinement. Many items that were generated contained simple semantic errors due to faults in the lexicon. Often these errors result from general noun or verb ambiguities that are not ambiguous in a particular domain such as electrical systems. For example, the verb *to type* appears in the English lexicon, but "type" is very unlikely to be used as a verb in an electrical context and is almost always a noun ("a type of pliers"), so the verb entry for "type" should be removed from the electrical systems lexicon. Another form of lexical error occurs when the lexicon does not correctly identify a noun as a mass noun (as opposed to a singular or plural noun). For example, in the electrical domain, the noun "power" is always a mass noun (circuits have "power" or "some power", never "a power" or "powers"). Once these minor adjustments were made to the lexicon, the item generation system was run again. At this point, 33 out of 110 items generated (roughly one in three) were viable (See Figure 2).

Source:	Solderless connectors, popularly called wire nuts, are used for connecting small-gauge and fixture wires.
Question:	what is the purpose of solderless connectors?
Key:	to connect fixture wires
Distractors:	a. to improve a grounding system b. to identify hot wires c. to indicate circuits

Source:	Round-nosed pliers are used for making screw loops and working in limited-space areas.
Question:	what is the purpose of round-nosed pliers?
Key:	to make screw loops
Distractors:	a. to anchor screws b. to identify hot wires c. to attach electrical devices

Figure 2. Sample Viable Items

Regarding the nonviable items, a small percentage reflected a level of ambiguity that require human understanding of the content and subject matter that a machine will not be able to process. For example, this phrase was taken from the electrical systems manual: "Two- and three-wire distribution systems, either direct current (DC) or single-phase AC, are widely used for lighting installations." In this case, the system may mistakenly interpret "lighting" as a verb, which would lead to the stem, "What is used to light installations?" Linguistically, this is a reasonable interpretation, but an experienced electrician would understand that "lighting installations" is used as a noun phrase.

Another important finding concerns the man hours spent to achieve results as seen in Figure 2. The specific hours spent to develop the initial grammars from the test-bed content (combat radio operator manual) cannot be accurately reported, as that task cannot be teased apart from other aspects of the research conducted during this time. However, the initial grammars required several months of work, while the minor adjustments needed to achieve the improved results for the electrical systems content took only a week.

Analysis

Based on the experience generating items from both the combat radio and the electrical systems manuals, it is clear that accurately extracting the important semantics from technical manuals and reliably generating viable assessment items requires a deep understanding of both the technical material and general world knowledge that automated systems do not have and are unlikely to have for many years to come. While highly accurate item generation is currently beyond the state of the art, high accuracy is not required for the technology to save instructors significant amounts of time. The percentage of viable items (e.g., 30% for the electrical systems manual) is large enough that it

should be faster to pick viable items from the generated list (including manually picking appropriate distractors from a list) than to manually generate items from scratch.

CHALLENGES AND IMPLICATIONS

The results demonstrate that an automated assessment system can potentially create viable items and distractors. The primary challenge will be efficiently translating this potential into complete, valid assessments. The algorithmic limitations are largely analogous to human limitations in understanding language. Poorly written text, ambiguous references, and lack of domain knowledge can confound the extraction technology and lead to nonsensical items and other issues. Unsurprisingly, the writing quality of technical manuals varies widely. Long paragraphs with heavy use of pronouns (e.g., this, it, etc.) or ambiguous subjects often require parsing text across multiple sentences. While this technology exists, it is at the bleeding edge of current computational linguistics research and is beyond the current scope of research, while it will certainly enhance the efficiency of this approach in the future. This challenge also underscores the importance of entity recognition and lexicon development to resolve ambiguous references (e.g., "lighting installations") and appropriately recognize named objects. While developing a domain-specific lexicon can be time-intensive, this could be offset if a lexicon were provided by a subject matter expert or included in source text or altogether avoided if a lexicon has already been developed for a given domain.

Lack of domain knowledge applies to both understanding radio technology (or electrical systems) in general and basic human understanding of objects and physics. This makes it challenging to eliminate nonsensical candidate items entirely without providing instructions regarding definitions, types, classes and so forth. To address the challenge of domain-specific knowledge, using the test-bed content, the team experimented with developing a domain ontology, a formal definition of relationships between elements. A radio is a piece of communication equipment, the combat radio is a type of radio, a Soldier is a person who would use a radio, and so forth. The team discovered that when this ontological information is available, it can greatly enhance clarity and reduce nonsensical item generation, but such information is costly to develop and maintain. This cost can be offset in domains where ontologies already exist, such as in the medical training domain, or offset over time through incremental additions to the ontology. Lack of general knowledge about the world is a challenge in all Artificial Intelligence (AI) research, not just in AIG.

Together, these challenges imply a stronger role in the near-term for a human review and computational linguist than originally hypothesized. While the prototype can generate large amounts of items and distractors, low-quality source text will potentially generate nonviable items and distractors that cannot be eliminated through algorithmic methods alone. While the current state of language technology prevents a completely "automated assessment," the process of reviewing and selecting viable items and distractors is straight-forward and efficient with human interaction. It also suggests that a more realistic goal of near-term research is not to eliminate the human, but to empower a human instructor with an item-generating "co-bot" that serves up item candidates for review and selection. It also implies a greater focus on elimination of nonviable options for human review, rather than automated generation of complete and ready-to-administer assessments. This realization led the team to focus on two key elements of human involvement requiring further research, which are addressed in the next two sections.

USER INTERFACE

One key element required for human involvement is the ability to easily generate and then refine the automated assessments through a system user interface. The research team consulted with a panel of assessment experts to design user interface mockups that could be developed into a fully functioning prototype in future research. The assessment experts provided input regarding features and functionality that would allow the user to first generate an assessment based on user specified criteria and then to further refine the assessment by accessing item stem, key, and distractor variants stored in the item database. The interface would provide mechanisms for filtering items based on user specified criteria and item and assessment metadata. The system would filter the available set of items to those that match the specified criteria and select the top items for a generated assessment available for the user to review. The user would then select items from the generated assessment to view item details and could also edit the item by selecting available item stem and key variants and recommended distractors. These sophisticated system features would facilitate the process of building high-quality assessments; however, the research team understands that such a user interface is better suited for instructors who are knowledgeable and experienced in assessment development. In the short term, a simple searchable database of automatically generated items could suffice for less experienced instructors and would still serve as a time-saver and enable more rapid development of effective assessments. Content

experts could review, vet, and select automatically generated items, which could then be exported into a compatible file format and subsequently imported into an existing program that houses an item bank.

ITEM DIFFICULTY METRICS

Another key element for human involvement is the ability to reduce the amount of analysis required to ensure appropriate item difficulty. Item difficulty refers to the extent to which test takers answer a test item correctly. Items that have a higher percentage of test takers answering correctly are easier than items that have a lower percentage of test takers answering correctly. Not only is it important to generate valid items based on training materials, it is also important to generate items at the correct level of reading difficulty, cognitive complexity, and at the appropriate level of learner instruction (e.g., beginner, intermediate, advanced). These factors all contribute to item difficulty. Estimating and controlling item difficulty is important for constructing effective assessments but is also critical for AIG. As explained, the automated technology can generate large numbers of item variations and response options that vary widely in difficulty. It would require a great deal of effort and time for a human to comb through all of the variants to select the best options. Item difficulty metrics that measure item variants' difficulty can help to reduce human requirements by automatically identifying item variants that better align with assessment requirements that the user specifies through the User Interface.

The team conducted an extensive literature review of the educational and industrial/organizational psychology literature as well as the AIG literature to identify various factors that have been shown or hypothesized to impact item difficulty. Based on the literature review, the team developed an item difficulty framework. The team is now in the process of generating several metrics based on that framework and will be conducting experiments to assess which metrics are most useful, particularly for narrowing the list of potential response options. It is anticipated that the system will still present the user with a number of distractor options for consideration, but the goal is to provide a more refined, better targeted list.

FUTURE RESEARCH

The semantic approach to automated assessment generation has great promise and is succeeding in generating many potentially viable items. Future research should focus on overcoming the engineering challenges necessary for efficient item generation. These challenges can be viewed in two different ways, both focusing ultimately on the Army instructor experience. The first will focus on validation that the candidate items and distractors are indeed viable for instructors and learners and predictive of future performance on job tasks. This would entail partnering directly with schoolhouses or units to evaluate generated items and tailor the output and user interface to their specific needs. More generalization studies, such as the team's effort with the electrical systems manual, should be conducted to assess how easily items can be generated for new courses or subject matter. Generalization studies that assess transferability across courses or subject matter could include scaling to broader educational contexts, such as other military schoolhouses, secondary education, or corporate learning and development.

The second challenge will focus on enhancing effectiveness at each stage of the assessment production pipeline, from text import and grammar extraction to selection of candidate items and distractors. The initial focus of these efforts will be reducing the number of nonviable items and distractors that the instructor must sift through. There are many opportunities to leverage other current research efforts ongoing in the governmental, academic, and commercial space. The most impactful computational linguistics advancements are likely to be parsing and extracting frames from across multiple sentences and improved recognition of technical names, both very active fields. At the other end of the pipeline, research efforts across the entire data science domain will aid in efficient and effective estimation of item difficulty.

REFERENCES

Ch, D.R., & Saha, S.K. (2018). Automatic multiple choice question generation from text: A survey. IEEE. *Transactions on Learning Technologies, 13(1);*14-25. https://doi.org/10.1109/TLT.2018.2889100.

Cubric, M. & Tosic, M. (2010). Towards automatic generation of e-assessment using semantic web technologies. In: *Proceedings of the 2010 International Computer Assisted Assessment Conference.* http://hdl.handle.net/2299/7785.

Flor, M., & Riordan, B. (2018). A semantic role-based approach to open-domain automatic question generation. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications,* pp. 254-263. https://www.aclweb.org/anthology/W18-0530.

Götz, F. M., Maertens, R., & Linden, S. (2021, June 14). Let the Algorithm Speak: How to Use Neural Networks for Automatic Item Generation in Psychological Scale Development. *PsyArXiv* Preprint.pdf. https://doi.org/10.31234/osf.io/m6s28.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: a comprehensive review. *Review of Educational Research*, 87, 1082-1116.

Halliday, M. A. K. (2003). *On Language and Linguistics*. Vol. 3 in the Collected Works of M.A.K. Halliday. Webster, J. (ed.). Continuum.

Halliday, M. A. K. (2007). *Language and Society*. Vol. 10 in the Collected Works of M.A.K. Halliday. Webster, J. (ed.). Continuum.

Halliday, M.A.K. & Matthiessen, C. (2014). *An Introduction to Functional Grammar*, 4th Edition. Routledge.

Kurdi, G., Leo, J., Matentzoglu, N., Parsia, B., Sattler, U., Al-Emari, S. (2019). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*; 121-204. https://doi.org/10.1007/s40593-019-00186-y.

Le, N.-T., Kojiri, T., & Pinkwart, N. (2014). Automatic question generation for educational applications–the state of art. van Do, T. et al (eds) *Advanced Computational Methods for Knowledge Engineering.* Cham: Springer, pp. 325–338.

Le, Q., Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning* (ICML), pp. 1188–1196.

Leo, J., Kurdi, G., Matentzoglu, N., Parsia, B., Forege, S., Donato, G., Dowling, W. (2019). *Ontology based generation of medical, multi-term MCQs. International Journal of Artificial Intelligence, in Education.* https://doi.org/10.1007/s40593-018-00172-w.

Liang, C., Yang, X., Dave, N., Wham, D., Pursel, B. & Giles, C.L. (2018). Distractor generation for multiple choice questions using learning to rank. *In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 284–290, New Orleans, LA. Association for Computational Linguistics. https://clgiles.ist.psu.edu/pubs/naacl18_bea.pdf

Long, R., Barnieu, J., Hyland, J. (2015), Development and Evaluation of Mobile Training Technologies. *In Proceedings of the Interservice/Industry Training, Simulation, and Education Conference* (I/ITSEC) 2015. Orlando, FL.

Mann, W., & Matthiessen, C. (1983). *Nigel: A Systemic Grammar for Text Generation* (No. ISI/RR-83-105). University of Southern California, Information Sciences Institute.

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. DOI: https://doi.org/10.48550/arXiv.1301.3781. arXiv preprint arXiv:1301.3781.

Papasalouros, A., Kanaris, K., & Kotis, K. (2008). Automatic generation of multiple choice questions from domain ontologies. In *Proceedings of the IADIS e-Learning Conference*, pp. 427–434, Amsterdam, Netherlands, July 2008. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.218.5149&rep=rep1&type=pdf

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C., and Scheffczyk, J. (2006). *FrameNet II: Extended theory and practice.* http://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf.

Spain, R., Harris Mulvaney, R., Cummings, P., Barnieu, J., Hyland, J., Lodato, M., & Zoellick, C. (2013). Enhancing Soldier-Centered Learning with Emerging Training Technologies and Integrated Assessments. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*. Orlando, FL.

Subramanian, S., Wang, T., Yuan, X., Zhang, S., Bengio, Y., & Trischler, A. (2017). Neural models for key phrase detection and question generation. *Machine Reading for Question Answering workshop at ACL 2018.* https://arxiv.org/pdf/1706.04560.pdf

Vinu, E.V., & Kumar P. S. (2015). A novel approach to generate MCQs from domain ontology: Considering DL semantics and open-world assumption. *Web Semantics: Science, Services and Agents on the World Wide* Web. http://dx.doi.org/10.1016/j.websem.2015.05.005.

Winograd, T. (1971). *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language.* MIT AI TR-235. Artificial Intelligence Laboratory, MIT.

Zhou, Q., Yang, N., Wei, F., Tan, C., Bao, H., & Zhou, M. (2017). Neural question generation from text: A preliminary study. *In National CCF Conference on Natural Language Processing and Chinese Computing,* pp. 662-671. Springer. https://arxiv.org/pdf/1704.01792.pdf.

Disclaimer: The research described herein was sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences, Department of the Army (Contract No. W911NF-20-C-0018). The views expressed in this presentation are those of the author and do not reflect the official policy or position of the Department of the Army, DOD, or the U.S. Government.