Improving Assessment with Text Mining

Hillary Fleenor, Rania Hodhod
Columbus State University
Columbus, Ga
fleenor_hillary@columbusstate.edu,
hodhod_rania@columbusstate.edu

Randy Brou
Army Research Institute
Ft. Benning, Ga
randy.j.brou.civ@mail.mil

ABSTRACT

Assessment is a key component of education across society. Regardless of whether the setting is academia, industry, military, or non-profit organizations, assessment is essential for gauging educational effectiveness, providing remediation to students, and informing policy and decision-making. However, the use of thorough assessments can be resource intensive. For example, instructors must devote time and effort into scoring/grading assessments. This can be especially costly when teaching complex skills that are not easily measured by convenient means such as multiple choice examinations (e.g., leadership, problem solving, critical thinking, communication). However, one can argue that these kinds of complex tasks are the ultimate goal of any educational system.

Computing holds great potential for reducing the burdens associated with assessment tools designed to measure complex skills. As a case in point, consider the Consequences test (Christensen, Merrifield & Guilford, 1953). It has been used to predict meaningful outcomes for military Officers, but the scoring of the test is extremely time-consuming as it requires test administrators to read and categorize test-taker-generated statements involving the outcomes of hypothetical scenarios. If the scoring of such statements could be automated, the test would become much easier to administer widely as the costs of the assessment would be drastically reduced. The challenge to implementing such a solution has been that computationally processing natural language, especially the kind of free form, conversational responses common in everyday life, is complicated. Nonetheless, tools already exist that show potential for utilization in assessment systems that necessarily use highly unstructured, free text input. In this paper, we discuss the use of open source Python libraries for assessing short answer, free form responses in the Consequences test. Using Latent Semantic Analysis, a well-established technique that has been around since 1988, we were able achieve human-computer response categorization interrater reliabilities comparable to human-human interrater reliabilities.

ABOUT THE AUTHORS

Hillary Fleenor is a lecturer in the TSYS School of Computer Science at Columbus State University. She is also a Senior Research Fellow with the Army Research Institute at Ft. Benning through the Consortium of Universities of the Washington DC Metropolitan Area, assisting with computing solutions to help improve soldier training. She has a master's degree in applied computer science and a master's degree in education. Her research areas include the use of artificial intelligence and text mining techniques to improve educational and training technologies.

Rania Hodhod, PhD is an Assistant Professor in TSYS School of Computer Science, Columbus State University. Rania's research interests span a range of areas, such as artificial intelligence, expert systems, serious games, interactive narrative and computational creativity. She has published over 40 refereed articles and 2 book chapters in these areas. Rania earned a PhD degree from University of York, UK. After joining Columbus State University in 2013, her research focused on educational games and developing a computational model for creativity and creativity assessment.

Randy Brou, PhD, Research Psychologist, Army Research Institute Fort Benning Research Unit, has seventeen years of experience in conducting applied research for the Department of Defense. He has led research projects for both the US Army and the US Navy focusing on training effectiveness and the measurement of individual and team attributes relevant for successful performance. Dr. Brou holds a Ph.D. in Applied Cognitive Science from Mississippi State University.

Improving Assessment with Text Mining

Hillary Fleenor, Rania Hodhod
Columbus State University
Columbus, Ga
fleenor_hillary@columbusstate.edu,
hodhod_rania@columbusstate.edu

Randy Brou
Army Research Institute
Ft. Benning, Ga
randy.j.brou.civ@mail.mil

INTRODUCTION

Assessment

Assessment is an invaluable tool for human progress; it is used to gauge learning, determine skill levels, and understand the human mind. Assessments are used by educators, medical professionals, psychologists, and government agents, to name a few. The purpose of assessment is to gain useful information that can be used to guide future actions. Good assessment can take us from blind guessing to knowledgeable decision making.

The information age has brought about changes in the workplace and in education that make assessment a fundamental necessity to determine individual needs and maximize efficiency. Assessments help provide guidance in a globally connected world with an ever-increasing amount of knowledge to learn and an ever-increasing number of options for individuals and organizations. It is a well-known fact that good assessment takes time. Assessments take time to create, time to administer, and time to score. This time demand increases for assessments that allow free text responses, as free text dramatically increases the number of degrees of freedom beyond selection types.

Assessments that allow free text responses can provide more detailed information to an assessor than is generally available in selection assessments such as multiple choice or true/false. However, free text assessments are particularly difficult to grade as they are not only time consuming, but are also more vulnerable to subjectivity in the assessment process. Recently computational statistics and machine learning have been explored for use in the automatic scoring of assessments that allow free text responses in hopes of reducing time burdens and reducing scorer bias.

The U.S. Army is one organization that utilizes assessments to inform a variety of human resources decisions. One assessment that is being studied for use at the present is the Consequences Measure. The Consequences Measure is a psychometric assessment that utilizes open-ended unstructured text responses. Researchers in the employ of the US Army have used the Consequences Measure in previous studies involving leadership (Mumford, Zaccaro, Harding, Fleishman, & Reiter-Palmon, 1993) and continue to use it to the present day. However, scoring the Consequences Measure is time consuming as each response must be carefully read, considered, and categorized. The set of responses provided by test-takers must also be read multiple times to check for duplicates. In addition, because of the subjective nature of free text scoring, the scorers must be extensively trained. The focus of this project is the exploration of methods for automating the scoring of this assessment to reduce the time input and human error involved in scoring.

Text Mining

Text Mining "can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analytical tools" (Feldman and Sanger 2007, p. 1). Text Mining can utilize approaches from statistics, knowledge discovery in data (KDD), machine learning, and data mining provided the unstructured text is preprocessed and transformed into an appropriate format (Hotho, Nurnberger and Paab 2005). In addition, text mining can employ methodologies and techniques from other computer science disciplines that handle natural language such as information retrieval, information extraction, natural language processing (NLP), and corpus-based computational linguistics.

Text mining systems rely on algorithms and statistical techniques that are used to discover patterns, analyze trends, or discover knowledge within a document set. "The three most common types of patterns encountered in text mining are distributions (and proportions), frequent and near frequent sets, and associations" (Feldman and Sanger 2007, p. 19).

Distributions are used to select subsets of collections that represent specific concepts by proportions or averages. Frequent and near frequent sets are sets of concepts represented in a document collection with co-occurrences above a selected threshold parameter (Feldman and Sanger 2007). Associations are directed relations between concepts or sets of concepts represented as rules (Feldman and Sanger 2007). Trend analysis focuses on changes in concept distributions over time. Concepts can be words, phrases, or patterns of word occurrences. Background knowledge can be used in a text mining system to aid in knowledge discovery.

There are a variety of specific statistical methodologies and algorithms available for accomplishing the above tasks. Selections are generally based on the goals of a particular text mining system. In addition to the broad category of knowledge discovery, text mining goals usually involve categorization, clustering, or information retrieval. Documents are preprocessed using techniques selected to prepare the text for the core mining operations to be performed. The processed document collection is then used in the chosen core mining operations by the user. Figure 1 shows a high-level representation of a general text mining architecture.

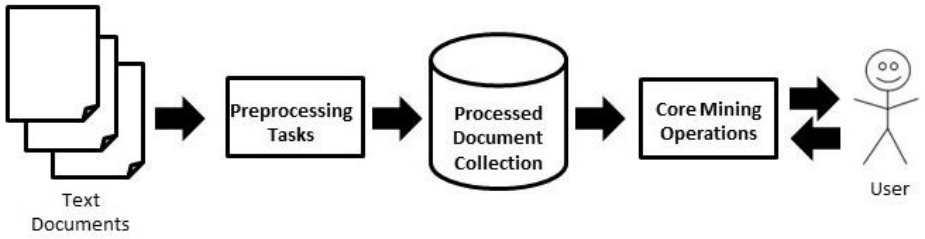

Figure 1. General Architecture of a Text Mining System

Text mining has been around for decades and there are numerous techniques that have already been developed along with tools to implement them. We made use of some of these established techniques and tools in this work, including Latent Semantic Analysis (LSA), and the Natural Language Toolkit (NLTK) implemented using Python 3 libraries.

The Consequences Measure

Psychometric assessments are a standard, scientific method used to measure individuals' mental capabilities and behavioral style. Psychometric tests are designed to measure candidates' suitability for a role based on the required personality characteristics and aptitude. This work explores the use of text mining and natural language processing to automatically score the psychometric assessment known as the Consequences Measure. The Consequences Measure was developed at the University of California as a part of the Aptitudes Research Project under the direction of J.P. Guilford (Guilford and Guilford 1980). The work started in the 1950's and was based on the Structure of Intellect Model (SOI) created by Guilford. Guilford's SOI Model was specifically designed as a frame of reference for intellectual abilities and has "served the heuristic function of generating hypotheses regarding new factors of intelligence" (Guilford 1967).

The Consequences Measure was developed as a measure of creative thinking capacity and focuses on the SOI operation of divergent production. Divergent production is the ability to generate multiple solutions to a problem (Guilford and Guilford 1980). The test primarily seeks to measure two important aspects of creativity: ideational fluency and originality. In the SOI Model, ideational fluency is identified with the Divergent Production of Semantic Units (DMU) while originality is identified with the Divergent Production of Semantic Transformations (DMT) (Guilford and Guilford 1980). DMU is the ability to come up with a variety of meaningful ideas in response to a question or problem. DMT is the ability to come up with meaningful and transformative ideas in response to a question or problem.

The version of the Consequences test currently utilized by the U.S. Army Research Institute (ARI) is a modification of the 1980 version that incorporates scoring methodology from an ARI study from 1997 (Dela Rosa, et al. 1997).

The test consists of a series of scenarios to which respondents are asked to supply a list of consequences. The four scenarios used by ARI are:

- What would be the results if the force of gravity was suddenly cut in half?
- What would be the results if human life continued on Earth without death?
- What would be the results if everyone suddenly lost the ability to read and write?
- What would be the results if it appeared certain that within three month the entire surface of the earth would be covered with water, except for a few of the highest mountain peaks?

Guilford's scoring system calls for scorers to rate each response as remote, obvious, irrelevant, or duplicate. Irrelevant and duplicate items are non-scoring. If a list of related items is given on multiple lines, the entire list is counted as one item. Scored items are obvious or remote. An obvious response is "one that appears to a direct result of the question, lacking features that would be commonly associated with the item, and a limited sense of the social, economic, or cultural consequences that the item would create" (Christensen, Merrifield, & Guilford, 1958, p. 1). Additionally, vague responses are also counted as obvious. A remote response is "one that indicates a consideration of changes that are more removed, temporally or geographically, or implies a specific substitute, a new system or some other fairly specific way of adjusting to the indicated changed situation" (Christensen, Merrifield, & Guilford, 1958, p. 1)

The Consequences Measure has been used in a number of factor analytic studies in military, educational, and other institutional settings. The score for obvious responses has consistently shown correlation with ideational fluency; the score for remote responses has been proven to be "a measure of 'originality' of the type having to do with remote associations or with revisions or transformations (Guilford and Guilford 1980). Studies have shown that these two scores are not correlated with one another.

Responses to the Consequences Measure are text entry in the form of an open-ended, unordered listing. The responses are generally sentence fragments rather than complete sentences and responses for a single question vary widely in content topics. Although research has been done on computer scoring of essays and single topic short answer questions (Klein, Kyrilov & Tokman, 2011, Mohler & Mihalcea, 2009, Pulman & Sukkarieh, 2005), we could find nothing in the literature on automatic scoring of free text sentence fragments that range widely across content topics. The lack of topic focus is a key difference. Short answers to assessment questions usually have a limited vocabulary set, e.g. for a short answer computer science question about round-robin scheduling, the vocabulary will be limited to terms that can be used to describe the round-robin scheduling process. Assessments that allow responses that range across topics require a larger vocabulary set. However, many assessments, especially those done outside of a formal academic setting, accept responses that range across topics so automatic scoring of free text sentence fragments that range widely across content topics has numerous applications.

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is "a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations" (Landauer, Foltz & Laham, 1998, p. 2). LSA has low computational processing requirements and eliminates the need for human powered rewriting of text in formal notation. In addition, LSA handles unique vocabulary use without having to program an understanding of these words into the system. By using LSA "a great deal of what is conveyed by a text can be extracted automatically using the tools of linear algebra. LSA provides not only that information explicit in the text, but also the underlying or latent meaning of the text" (Landauer, McNamara, Dennis & Kintsch 2007, p. x-xi). In addition, LSA is available in many pre-existing tools as it has been used for several different applications, including information retrieval, for decades. Table 1 shows that out of eight studies that used text-mining for assessing short answer responses, six used Latent Semantic Analysis (LSA) either alone or in conjunction with other methods.

Table 1. Text Mining for Assessment of Short Answer Responses

Authors	Text Mining	Methods Used	Research Goals	Results
(Wiemer-Hastings, Wiemer-Hastings & Graesser, 1999)	Categorization	LSA, cosine similarity	Improve an Intelligent tutoring system with automatic scoring	0.49 correlation with human raters compared with 0.51 between human raters
(Pulman & Sukkarieh, 2005)	Information Extraction, Clustering	Inductive Logic Programming, Decision Tree Learning, Naïve Bayesian Learning	Investigate computational linguistics for short answer marking	IE with pattern matching outperformed machine learning.
(Mohler & Mihalcea, 2009)	Categorization, Clustering	LSA, ESA, TF-IDF, WordNet shortest path,	Explore techniques for automatic short answer grading	Best performance: LSA with a large topic specific corpus
(Klein, Kyrilov & Tokman, 2011)	Categorization	LSA, TF-IDF, cosine similarity	Use LSA to automatically grade short answer CS questions	Pearson correlation of 0.8 or higher on all but one question
(Mohler, Bunescu & Mihalcea, 2011)	Categorization	Dependency graph alignment, LSA, isotonic regression	Use machine learning to improve lexical semantic similarity	Results varied based on the performance measure used
(Basu, Jacobs & Vanderwende, 2013)	Clustering	Similarity measures using NLP, TF-IDF, LSA; logistic regression; k-metoids algorithm, LDA	Using machine learning to improve an augmented LSA baseline	Grading accuracy improved from 82.5% to 92.5%
(Krithika & Narayanan, 2015)	Categorization Clustering	K-means, linear and isotonic regression, lexical similarity, structure features, dependency graphs	Build a scoring system in MS Azure that uses a combination of text mining and NLP techniques	All 76 answers tested were graded correctly
(Zehner, Salzer & Goldhammer, 2015)	Clustering	NLP, hierarchical clustering with LSA, ESA and distance measures using different agglomeration methods	Investigate the use of clustering to automatically code short answer responses.	The system reached high percentages of agreement on test data across all systems (from 76% to 98%)

Research has shown that Natural Language Processing (NLP) techniques can improve the LSA results by reducing words to base semantic units (Basu, Jacobs & Vanderwende, 2013). The following two NLP techniques are used in this work:

- **Stemming**: remove endings from words such as those that denote verb tense or plurality.
- **Lemmatization:** map words to their base word such as verb forms to their infinitive tense.

Research also shows that term weighting can also improve the results of LSA (Basu, Jacobs & Vanderwende, 2013). A commonly used weighting scheme is Term Frequency-Inverse Document Frequency (TF-IDF); the "TF" normalizes term occurrence by taking into account document size and the "IDF" eliminates words that occur frequently among all documents and are, thus, unlikely to offer meaningful data.

Interrater Reliability

In this work, we use inter-rater reliability (IRR) to determine the effectiveness of the automatic scoring methods. IRR "is the technical term used to describe how closely raters agree with each other" (Haley, et al. 2009, p. 85), i.e. a numerical measure/estimate of the degree of agreement between raters. IRR helps to ensure that assessment is consistent. It is important in assessment for scoring to be consistent in order for score data to be useful. Consistent scoring is a requirement to be able to make comparisons between individuals as well as between time points for a single individual.

Studies have shown that IRR between humans varies widely due to marker bias, inconsistency, etc. The authors of (Haley, et al. 2009) found IRR values that ranged from 0.15 to 0.97 for pairs of five expert scorers on 18 different questions from an introductory computing course. The authors of (Basu, Jacobs and Vanderwende 2013) found IRR values between three scorers for 10 questions from a citizenship exam ranged from 0.45 to 0.99.

It seems reasonable, as argued by (Haley, et al. 2009, p. 83), that an automatic scoring system is "*good enough* if it agrees with human markers as well as human markers agree with each other". Therefore, our goal is to develop a computational method or methods that achieve adequate levels of IRR with humans in scoring Consequences responses as compared to the IRR that humans have with one another.

There are different approaches to calculating IRR both conceptually as well as mathematically. These can be grouped into one of three general categories depending on the underlying goals of the analysis: consensus estimates, consistency estimates, or measurement estimates (Stemler 2004). These are defined as follows from (Stemler 2004):

- Consensus estimates – "based upon the assumption that reasonable observers should be able to come to exact agreement about how to apply the various levels of a scoring rubric" (p. 2-3).
- Consistency estimates – "based upon the assumption that it is not really necessary for two judges to share a common meaning of the rating scale so long as each judge is consistent in classifying the phenomenon according to his or her own definition of the scale" (p.6).
- Measurement estimates – "based upon the assumption that one should use all of the information available from all judges (including discrepant ratings) when attempting to create a summary score for each respondent" (p.8).

In this work, we use consensus estimates since these techniques are well suited to analysis in which variables being rated represent qualitatively different categories (Stemler 2004) as is the case with the obvious and remote categories of the Consequences Measure. The other two estimates are more useful for performance summaries. Common methods for calculating consensus estimates include: the simple percent-agreement figure, modification of the simple percent agreement scale to include adjacent categories, Cohen's kappa statistic, and Fleiss' kappa statistic. In this work, simple percent agreement and Fleiss' kappa are used. Simple percent without modification is used to compare pairs of judges, as this was used by the original scorers of the data set. Intermediate steps in the scoring process were not preserved so Cohen's kappa could not be calculated for the original scoring. Simple percent without modification is used throughout for consistency. For the multi-topic method where scoring is redone, Fleiss' kappa is calculated in addition to simple percent without modification.

It is important to note that the use of Pearson's coefficient is often used in automatic scoring literature because the assessments being researched often involve scoring on a graded scale, e.g. from A to F. Pearson's coefficient is

regarded as a measure of consistency which is very useful in analyzing automatic scoring of a scaled assessment, i.e. performance summary grading. However, it is not useful for the Consequences Measure since obvious and remote are discrete categories and the meaning of each is germane to the results.

METHODS

In this work, we chose to take a categorization approach to scoring, selecting categories and training a system to assign responses to the appropriate category. LSA is used on the training data and responses to reduce each item to a pattern of frequently occurring terms. Then cosine similarity is used as the measure for classification of each response in the test data to a category comprised of training data. Cosine similarity was chosen because of its ease of implementation and its independence of document length (Huang, 2008). Natural Language Processing (NLP) techniques (Lemmatization, Stemming) and TF-IDF weighting were explored to see if they would impact results. We tried two different approaches to categorization. The first used only two categories: Obvious and Remote. In the second approach, we created a set of customized categories for each question and assigned each category as obvious or remote.

Data Collection

The data used for both training and testing was collected prior to this work as a part of the administration of an assessment to collect information that will be used to study the leadership qualities of Army Officers. The test was administered to Officers attending Officer leadership courses (Infantry Basic Officer Leadership Course (IBOLC), Armor Basic Officer Leadership Course (ABOLC), Officer Candidate School (OCS), or the Maneuver Captain's Career Course (MCCC)). The exam was computerized and delivered via an application.

The responses used in this work were collected from Captains (CPTs) and Lieutenants (LTs) The responses for each question were combined. For the two-topic approach, the responses were then separated into two documents: one document containing responses human scored as obvious and the other document containing responses human scored as remote. For the multi-topic approach the responses were separated into multiple documents (one document for each category). The responses were checked for spelling errors and corrected. Eighty percent of each category were chosen randomly for the training data sets. The remaining 20% was reserved for testing the system.

System Development

Both systems were developed in Python 3, a programming language well suited for text-mining applications. We utilized Gensim, an open-source toolkit for vector space modeling and topic modeling. Gensim uses the NumPy and SciPy libraries (Radim Rehurek 2011). The Natural Language toolkit (NLTK), an open-source platform for building Python programs to work with human language data, was also used in the system (Project 2016). Similarity measures were gathered for all variations including the null of stemming or lemmatization, and using TF-IDF weighting. See Figure 2 and Figure 3 for system architecture.

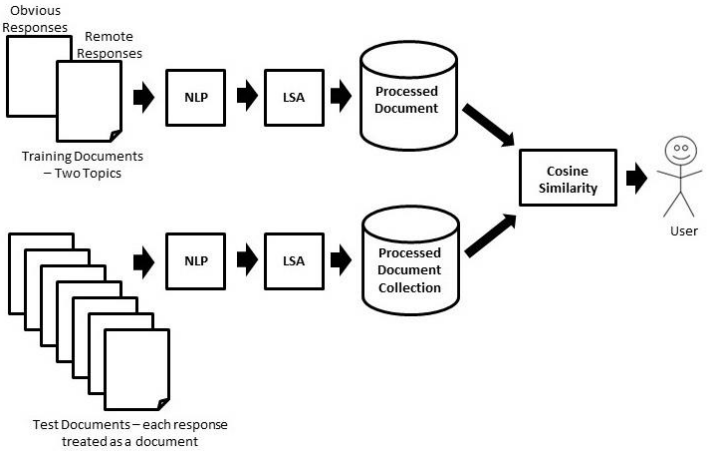

Figure 2. General architecture of two-topic scoring system

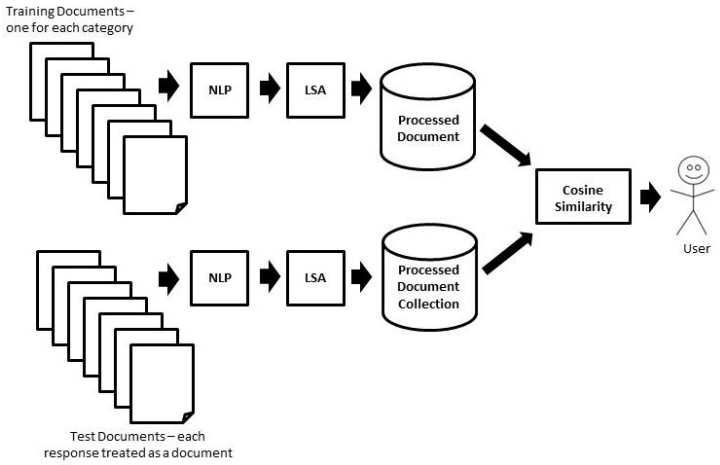

Figure 3. General architecture of multi-topic scoring system

Human scoring and Category Creation

For the two-topic approach (obvious and remote), a subset of the collected data was used that had been previously scored by a research psychologist at ARI and Consortium Fellow graduate students. This data set included a group of Lieutenants (LTs) from an IBOLC class of 116 officers and a small group of 30 Captains (CPTs) from the MCCC. To improve consistency in scoring, the scorers developed extensive scoring keys by modifying and adding on to the original scoring guides in (Guilford and Guilford 1980). The IRR for the scoring of LTs responses using the simple percent agreement figure was 91%. The IRR for the scoring of the CPTs responses using the same calculation was 83%. For responses where disagreement occurred, a discussion ensued to decide the final score. It is important to note that the high IRR values here are due to the extensive time put into creating the scoring guide. The data itself was used to create the scoring guide, so differences were resolved before the actual scoring. IRR is lower when scoring is done independently as is seen in the rescoring done for the multi-topic approach.

Research done by Mohler and Mihalcea achieved a nearly 3.5% increase in correlation using a topic specific corpus for training LSA over a generalized corpus (Mohler and Mihalcea 2009). Based on this, for the multi-topic approach, the same data set was used but the decision was made to separate training responses for each question into a number of categories in a way that each topic would be either obvious or remote. Using this method, test responses would be labeled as obvious or remote depending on which category they have the highest similarity with.

The use of a collaboratively developed scoring sheet has been used by prior researchers using the Consequences Measure. However, this results in an inflated inter-rater reliability score since disagreements are resolved before actual scoring. To get a clearer understanding of the variation in human scoring, rescoring of the data as obvious or remote was undertaken as a part of the categorization process. Raters were not provided with a scoring guide nor was a scoring guide developed as part of the scoring process.

A small subset of the data was used to develop an initial set of categories for each scenario. A Microsoft Access application was developed to guide the categorization and scoring processes. A total of 9 individuals from ARI, five research psychologists and four consortium students, participated in the scoring and category assignment of responses. As previously mentioned, scoring guides were not given to the raters. Instead, they were asked to carefully consider the definitions of obvious, remote, and irrelevant as given in the Consequences Measure scoring instructions. Raters were also asked to indicate the confidence with which they felt a response was obvious or remote; fully confident, confident, or slightly confident. Each question had three scorers. The results for the four questions are shown in Table 2. Both simple percent comparisons between pairs of scorers and Fleiss kappa values are shown. A Fleiss value between 0.01 and 0.20 is generally accepted to indicate slight agreement. For two scorers, 75% is considered minimal

agreement for simple percent agreement. These IRR values are significantly lower than the values achieved by developing a scoring sheet.

Table 2: Human scoring IRR

	Simple % 1 / 2	Simple % 1/ 3	Simple % 2 / 3	Overall	Fleiss kappa
Gravity	0.397	0.591	0.558	0.313	0.153
No Death	0.567	0.354	0.377	0.239	0.107
No Read/Write	0.557	0.382	0.506	0.277	0.095
Water	0.463	0.569	0.502	0.313	0.138

The training data from each scenario was used to finalize the categories. The goal was to create a set of categories where each category contains a minimum of 30 related responses of which the large majority scored as either obvious or remote. In this way, a response in the test data classified to a particular category would be scored as either remote or obvious (or irrelevant) based on the remote or obvious (or irrelevant) designation of that category. Inconsistencies in scoring semantically similar responses were resolved using the majority score. Some categories had to be split or combined. Confidence levels were very useful in this process. A few categories had to be designated as not computer scorable due to a large number of unresolvable conflicts.

In addition to the challenges of designating a category as either obvious or remote, some categories contained only one or two responses while others had well over 50. It was necessary to augment each category that contained less than 30 responses with additional responses to ensure a large enough vocabulary set. First, other data sets that had not been used were mined for responses that would fit a category. If this was not enough to reach 30 responses, then responses were created that were grammatically similar to existing responses using synonyms. Voyant, a set of text mining visualization tools (Sinclair and Rockwell 2016), was useful in this process as were WordNet (Fellbaum and Tengi 2016) and ConceptNet 5 (Speer and Havasi 2016), two online semantic networks. WordNet and ConceptNet 5 were especially useful in identifying synonyms and related concepts to use in searching other officer data sets for appropriate responses as well as in identifying useful synonym replacements.

RESULTS

Two Topic Categorization

Test data was run and the results compared to the human scoring. Comparisons were made with and without threshold values for similarity. Using no threshold performed significantly better than 5%, 10%, or 20% thresholds. The numbers of responses that agreed and disagreed with the human scoring were collected as well as the number of responses for which the system was unable to categorize as either obvious or remote. This was done for each of the test conditions. IRR was calculated using the simple percent agreement figure. Table 3 below show the results from the test data.

Table 3: IRR results. C is the control with no stemming, lemmatization, or TF-IDF. The S represents stemming, L represents lemmatization, and T represents TF-IDF.

	C	T	S	ST	L	LT
Gravity	0.762	0.826	0.740	***0.857***	0.759	0.820
No Death	0.827	0.836	***0.873***	0.857	0.860	0.859
No Read/Write	0.711	0.750	0.716	0.767	0.693	***0.775***
Water	0.664	0.825	0.682	***0.861***	0.683	0.820

As can be observed from Table 3, each question had a condition that achieved 75% or better IRR, although the test conditions achieving this varied by question. The No Death question had greater than 75% for each condition with the highest being 87.3% for stemming only. None of the test conditions for any of the four scenarios were able to achieve the level of IRR that highly trained human scorers using a well-developed scoring guide reached with the LTs responses (91%). However, at least one condition in three of the scenarios reached the lower IRR achieved with the CPTs responses (83%). Unfortunately, it was not consistently the same method and in the cases where a higher IRR resulted from using TF-IDF, the proportion of responses that were not machine scorable went up significantly. It is

again important to note that using a scoring guide developed from the actual data results in significantly higher IRR scores than independent scoring.

Multi-Topic Categorization

IRR was first calculated using Fleiss kappa values for the three human raters and the computer. As can be seen in Table 4, the resulting values range from 6.8% to 17.7% agreement above agreement that would be achieved through random chance alone. These are low, but comparably low to the Fleiss kappa values for the three human raters reported in Table 2 above.

Table 4: Fleiss kappa IRR results. C is the control with no stemming, lemmatization, or TF-IDF. The S represents stemming, L represents lemmatization, and T represents TF-IDF.

	C	T	S	ST	L	LT
Gravity	0.112	0.16	0.133	0.167	0.137	0.170
No Death	0.109	0.096	0.118	0.112	0.121	0.105
No Read/Write	0.073	0.085	0.089	0.085	0.084	0.068
Water	0.151	0.154	0.139	0.168	0.155	0.177

To gauge the effectiveness of the categories, IRR was also calculated using simple percent. However, instead of comparing the computer-generated rating to either an average rating or individual human rating, it was compared to the rating of the mode of the human selected category for each response. That is, a response was assigned the designation (obvious or remote) of the most frequently human selected category for that response. In cases where each scorer selected a different category, the response was marked as unscorable. As can be seen in Table 8, these range from 47.8% agreement to 71% agreement.

Table 5: Simple percent IRR results. C is the control with no stemming, lemmatization, or TF-IDF. The S represents stemming, L represents lemmatization, and T represents TF-IDF.

	C	T	S	ST	L	LT
Gravity	0.624	0.682	0.611	0.713	0.631	0.701
No Death	0.613	0.597	0.64	0.618	0.634	0.624
No Read/Write	0.554	0.582	0.614	0.592	0.571	0.603
Water	0.478	0.516	0.522	0.535	0.497	0.529

Although Table 4 shows only slight agreement above random chance and Table 5 shows agreement values near 50%, these are comparable to the low human IRR values achieved with independent scoring.

ANALYSIS OF RESULTS

Categorization performs nearly as well as human scoring. The two-topic method requires the least amount of user input and therefore is the preferred method. Although TF-IDF improved results overall, it resulted in a larger number of responses that could not be categorized. This is undesirable due to the large volume of responses that will need to be scored by hand. Stemming and lemmatization did not improve results significantly (nor consistently).

Although the multi-topic system seems like a promising tactic to narrow the vocabulary set for comparison, the open nature of the Consequences Measure made it challenging to determine a representative set of topics for each scenario. The results did not show a significant improvement over the two-topic system.

There are several threats to validity in this study. One problem is, although the rating software developed to collect data for the multi-topic categorization and the independent scoring included the capability for the user to go back and change responses if their scoring evolved over time, there were no instructions asking them to do so. This resulted in some inconsistency in the independent scoring. For example, one scorer of the Water scenario rated the response "colonize the moon" as confident remote, but "create a moon base" as confident obvious. Another possible source of error is that it was necessary to make a judgement call on categorizing responses with no topic consensus. Furthermore, there was no way to ensure the Consequences Measure was taken seriously by Army Officers being assessed and we occasionally found responses that indicated otherwise. For example, one participant entered the letter "a" on every

line. Human scorers reported struggling with scoring. Even when agreement is reached, it is always possible for two or three people to agree but both (or all) be incorrect.

DISCUSSION

Automatic scoring of free text short phrases that may or may not be in complete sentence form and that cover a wide variety of subjects is extremely difficult. Humans have the ability to make inferences about such a response by drawing on a wide range of personal experiences from memory about the scenario as well as the subject addressed in the response. Once having retrieved information they are then able to determine what is relevant to help them understand what might be meant by the response and then decide if it represents an obvious or remote response. The computer has to work with what it is given; it can take a great deal of processing time searching through information. A brief phrase by itself contains very little information to use to automatically provide the system with appropriate relevant context. LSA finds latent patterns that can sometimes be a useful substitute for context. LSA performs best with a large amount of text, which is lacking for this problem.

However, despite the brevity of the responses, LSA performed as well as human scorers for both methods we used, categorization with two-topics and categorization with multiple topics. The results from this work show that categorization using LSA gives acceptable results for use in automatic scoring of the Consequences Measure. Using multiple topics did not seem to improve results enough to justify the time input for creating and refining categories. The results of this work will inform future projects at ARI including computerized scenario based tool aimed at assessing, and potentially training, qualities needed for effective soldiers including leadership, problem solving, and communication.

REFERENCES

Basu, Sumit, Chuck Jacobs, and Lucy Vanderwende. 2013. "Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading." Transactions of the Association for Computational Linguistics 1: 391-402.

Dela Rosa, Michelle R., Deirdre J. Knapp, Brian D. Katz, and Stephanie C. Payne. 1997. Scoring System Improvements to Three Leadership Predictors. Alexandria, Va: Human Resources Research Organization.

Feldman, Ronen, and James Sanger. 2007. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. New York: Cambridge University Press.

Fellbaum, Christiane, and Randee Tengi. 2016. WordNet: A lexical database for English. May 8. Accessed 2016. https://wordnet.princeton.edu/.

Guilford, J. P. 1967. The Nature of Human Intelligence. New York: McGraw-Hill.

Guilford, J. P., and J. S. Guilford. 1980. Consequences Sampler Set: Manuel of Instructions and Interpretations. Palo Alto: Mind Garden.

Haley, Debra, Pete Thomas, Marian Petre, and Anne De Roeck. 2009. "Human fallibility: how well do human markers agree?" Proceedings of the Eleventh Australasian Conference on Computing Education. Wellington.

Hotho, Andreas, Andreas Nurnberger, and Gerhard Paab. 2005. "A Brief Survey of Text Mining." LDV Forum 20 (1): 19-62.

Huang, A. 2008. Similarity measures for text document clustering. In “Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008),” Christchurch, New Zealand: 49-56.

Klein, Richard, Angelo Kyrilov, and Mayya Tokman. 2011. "Automated Assessment of Short Free-Text Responses in Computer Science using Latent Semantic Analysis." Proceedings of the 16th annual joint conference on innovation and technology in computer science education. Darmstadt.

Krithika, R, and Jayasree Narayanan. 2015. "Learning to Grade Short Answers using Machine Learning Techniques." Proceedings of the Third International Symposium on Women in Computing and Informatics. Kochi.

Landauer, Thomas K., and Susan T. Dumais. 1997. "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge." Psychological Review 211-240.

Landauer, Thomas K., Danielle S. McNamara, Simon Dennis, and Walter Kintsch. 2007. Handbook of Latent Semantic Analysis. Mahwah: Lawrence Erlbaum Associates.

Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. 1998. "An Introduction to Latent Semantic Analysis." Discourse Processes 25: 259-284.

Mohler, Michael, and Rada Mihalcea. 2009. "Text-to-text similarity for automatic short answer grading." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.
Mohler, Michael, Razvan Bunescu, and Rada Mihalcea. 2011. "Learning to Grade Short Answer Responses using Semantic Similarity Measures and Dependency Graph Alignments." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.

Project, NLTK. 2016. Natural Language Toolkit. April 9. Accessed September 21, 2015. http://www.nltk.org/.

Pulman, Stephen G, and Jana Z. Sukkarieh. 2005. "Automatic short answer marking." Proceedings of the second workshop on Building Educational Applications using NLP. Ann Arbor.

Radim Rehurek, R. 2011. "Scalability of Semantic Analysis in Natural Language Processing." Brno: Masaryk University.

Sinclair, Stefan, and Geoffrey Rockwell. 2016. Voyant Tools. Accessed May 2016. http://voyant-tools.org/.

Speer, Rob, and Catherine Havasi. 2016. ConceptNet 5. Accessed 2016. http://conceptnet5.media.mit.edu/.

Stemler, Steven E. 2004. "A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability." Practical Assessment, Research & Evaluation 9 (4): 1-19.

Wiemer-Hastings, Peter K., and A. Graesser. 2004. "Latent Semantic Analysis." Proceedings of the 16th international joint conference on artificial intelligence. 1-14.

Wiemer-Hastings, Peter, Katja Wiemer-Hastings, and Arthur Graesser. 1999. Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. Vol. 99, in Artificial Intelligence in Education, vol. 99. Amsterdam: IOS Press.

Zehner, Fabian, Christine Salzer, and Frank Goldhammer. 2015. "Automatic Coding of Short Text Responses via Clustering in Educational Assessment." Educational and Psychological Measurement 1-24.

The research described herein was sponsored by the Army Research Institute for the Behavioral and Social Sciences, Department of the Army. Opinions, interpretations, conclusions, and recommendations are those of the author(s) and are not necessarily endorsed by the U.S. Army.