Investigating the use of vector analysis to assess students' understanding

Andrea M. Goncher[a] and Wageeh Boles[b]
Charles Sturt University[a], Queensland University of Technology[b]
Corresponding Author Email: agoncher@csu.edu.au

CONTEXT
Many methods have been proposed in the study of linguistics for the representation of words and sentences. Most classical methods are symbolic and consist in things like dictionaries, thesauri, ontologies and syntax trees. Another approach is to represent words and sentences via the use of high dimensional vectors, which capture the distributional statistics of words and sentences. One application of representing words as vectors is to automatically evaluate text, which can further be applied to the assessment of students' text-based answers.

PURPOSE
This study investigated approaches to automatically analyse student responses to questions in the signal processing domain.

APPROACH
We investigated vector analysis approaches to capture various semantic and syntactic features of words, such that these representations can be compared and scored in a graded fashion, as distinct to simply true/false or same/different. The approaches used in this study can be trained in a semi-supervised fashion, where minimal human input is typically required.

RESULTS
The data investigated in this study consisted of student responses to short-answer questions in text form with associated metadata indicating the correctness for answers. Difficulties encountered when automatically assessing student short answers, either for correctness or knowledge gaps, were a) variations in vocabulary b) variations in grammatical structures c) precisely determining when specific concepts occur and don't occur, and d) relevant concept modifiers that may alter the assessment of the short answer. One element—important for addressing these difficulties— is how words and sentences are represented in short-answer question responses.

CONCLUSIONS
The study described in this paper focused on vector space representations for text. We recommend the development an agile methodology to be employed so that regular outputs be produced and sent for comment, which can then be used to inform further work. We suggest the best approach is to make use of a combination of methods including the many classical Natural Language Processing (NLP) techniques such as part of speech (POS) tagging, and phrase chunking.

KEYWORDS
text analysis, conceptual understanding, machine learning.

Introduction

Many methods have been proposed in the study of linguistics for the representation of words and sentences. There are a number of classical Natural Language Processing (NLP) techniques, such as Part Of Speech (POS) tagging and Phrase Chunking, that have been used for processing textual data (Bates, 1995). Vector space models have been successful in the domain of information retrieval, which inspired its application to semantic tasks in NLP (Turney & Patel, 2010).

A primary division in these methods is between symbolic and connectionist (subsymbolic) approaches. Most classical methods are symbolic, and are used in tools such as dictionaries, thesauri, ontologies and syntax trees. Another approach is to represent words and sentences through the use of high dimensional vectors, which capture the statistical distribution of words and sentences. A related method is the method of constructing word embeddings. The idea of representing words and sentences as a point (or vector) in space helps to visualize the representation that points which are closer together are considered more semantically similar, and conversely, points that are farther apart are less similar (Turney,& Pantel, 2010). The similarity between words is based on the context in which they occur (Jurafsky & Martin, 2009), so words that have different meanings in one context will not be considered semantically similar.

In this paper, we explore the feasibility of vector-based analysis approaches to evaluate short answer (text-based) marking in an engineering context. The ability to computationally assess and evaluate short answer responses for conceptual understanding in engineering disciplines is important in understanding the structural differences between students' submitted answers and the identified correct answers to establish some level of accuracy of submitted responses. This paper is framed by the research questions: *1) what are the current methods and limitations of applying vector analysis to text*, and 2) *to what extent can textual data (words and sentences) be leveraged to improve upon existing approaches to short-answer evaluation?*

An outcome of this study is to inform and develop the ways in which an assessor's workload may be reduced, and inform the instructor (and student) of the degree that the submitted answers match correct answers. The primary difference between the requirements of an automated assessment system, and the requirements of automated systems identified in the literature, is the domain of the assessment, specifically signal processing. Signal processing (and other engineering disciplines), predominately uses mathematical symbols and equations to represent concepts.

Rationale for representing words as vectors

Multiple choice and numerical-only answers, provide binary assessment of right or wrong, and it is not easy to evaluate the student's understanding of the concept being testing. Specific multiple choice tests exist that purposely design the distractor responses (incorrect possible selections) to indicate the misconception associated with that response. The multiple choice selection alone does not provide information regarding the degree of understanding, or if the selection was chosen as a "guess" (Goncher, Jayalath, Boles, 2016). Test questions that are more likely to capture and uncover a students' understanding involve more explanatory answers, which are typically captured through written (or spoken) answers. Instructors can be hesitant to include problems in assessments that require multiple solution steps or long written explanations because of the time required to mark and effectively evaluate the response (Aggarwal, Srikant, Shashidhar, 2013). Representing words and sentences via the use of high dimensional vectors, which capture the

distributional statistics of words and sentences is informative to developing a system that can identify whether new input text is related to a set of identified text.

Representing words and sentences via the use of high dimensional vectors has recently increased in popularity (Turney & Patel, 2010). A related method is the method of constructing "word embeddings" (Mikolov et al. 2013) and has become popular in recent years. These approaches capture various semantic and syntactic features of words such that the representations can be compared and scored in a graded fashion, as distinct to simply true/false or same/different. They are also trained in an semi-supervised fashion, where no or little human input is typically required. This is an important feature, especially for the development of automated assessment systems.

An automatic short, text-based answer marking system could automatically evaluate an answer provided by a student, usually by comparing it to one or more expert (correct) answers. This approach would be different from a related method of keyword or paraphrase detection, because we would want the requirement of the system to evaluate some level of understanding (of the student) from the text, rather than assign a correct/ incorrect classification.

The vector space models (VSMs) represent words in a continuous vector space, where semantically similar words are mapped to nearby words that appear in the same contexts. Being able to compare given words, or text, to words in similar contexts would provide a more informative evaluation of whether the given student answer is similar to a "correct" response, which is derived from the response an expert would provide.

Study Design and Methods

An effective way to test the developed software is to also develop a number of "unit tests," which test different features of the software. For example, a list of short-answers could be constructed that would test the software's ability to correctly determine the existence of mathematical symbols or expressions. The same could be done for detecting guesses and other inconsistencies in their answers and understanding. The ultimate benefit of the software, or automated system, lies in the adaptability of the system to evaluate text-based responses to different questions without creating a new evaluation tool each time.

The following sections describe the data we collected from students to use in our investigation of techniques, and the various methods we reviewed and evaluated based on the collected data set.

Descrition of the Data

To investigate approaches for analysing text, we utilised a data set of questions from a conceptually-based test in electrical engineering, i.e. the Signals and Systems Concept Inventory (Wage, Buck, Cameron, Welch, 2005). The concept inventory test was administered as part of a digital communications course, and the text-based responses were provided by a class of undergraduate students. The student answers were collected via an online test format in order to obtain data that was easier to pre-process.

The data consisted of students' short-answers in text form with associated metadata, which included various forms of answers provided by an expert in the signal processing domain, to indicate the correctness of answers. Other metadata includes some exemplary responses constructed by experts. Many answers include mathematic expressions and references to diagrams. Answers provided by the students and experts generally consisted of one or several phrases.

Table 1 shows the correct answer provided by the expert and example student answers. The "incorrect" or "correct" label for the example student answers corresponds to the multiple-choice selection associated with the question. Ninety-six students were enrolled in the class, and 82 students submitted answers to 15 questions from the Signals and Systems Concept Inventory (Wage et al.,2005). The data set we utlised consisted of a total of 1,230 student text responses and the corresponding multiple choice selection answer set. In addition, the experts rated additional student answers that were used as part of the metadata to build a model of correct answers for each question.

Table 1. Example items from the data set

	DT-SSCI Question category, correct/ expert answer, and example student answers
Question: *Correct answer:*	3. (time reversal) *p[-n] is the time reversed signal of p[n]. Therefore, p[2-n] can be obtained by shifting p[-n] by two samples to the right.*
Student answer 1: Student answer 2:	"Reversed, shifted two units to the right." (incorrect) "signal is flipped and shifted to the left by 2" (incorrect)
Question: *Correct answer*:	8. (sampling) The sampling frequency 5 Hz, is greater than twice the frequency of the signal, which is 2 Hz.
Student answer 1: Student answer 2:	"To sample at 5Hz the signal needs to occur at 2.5Hz or lower" (correct) "because it hasnt been shifted and takes 0.2 seconds to complete one full wave" (incorrect)

Vector Space Models

Lexical Vectors

It is also possible to build vectors based on the surface form of words, which enables the comparison of words based on the string of characters of which they are constituted. The lexical vectors encode character *n-grams* within words. For example, if you wanted to break up the text into configurable sizes, the word "dogs" for a 1-gram and 2-gram would be broken up into "d", "o", "g", "s", "do", "og", and "gs." N-grams within words are particularly useful for identifying tokens, such as mathematical equations and imprecise matching.

Clustering

Clustering can be performed on vectors constructed from token sequences of specific length, sentences, or complete short answers. Clustering is useful for identifying recurring lexical and semantic patterns in short answers. Clustering can be used to identify important features. Clustering can be performed on: a) all available text data, b) all text for a specific question, across all students, and c) all text for a student, across all questions. In clustering, the basic idea is to group phrases into different groups based on a suitable similarity measure.

Table 2 provides an example of a cluster of fragments extracted from student responses in our data, using lexical vectors. The columns together form a single cluster, and capture the mathematical expressions from the data set.

Table 2. Clustering example

Cluster: Mathematical expressions	
5: p[n-2]	n=2: r[n] -
1: p[n-2] =	For n=2: r[n]
= 2: p[n-2]	x[n] -> y[n]
r[n-2] = r[(1)]	x[n] -> y[n]
r[n-2] = r[(2)]	p[n] *convolution p[n]
r[n-2] = r[(-1)]	*convolution p[n] ->
r[n-2] = r[(0)]	= p[n] *convolution
For n=0: r[n]	the function r[n-2]
r[n] - r[n-2]	and r[n]-r[n-2]=1 when
r[n-2] = r[(1)]	for p[n-2] has
For n=2: r[n]	p[n-2] is just

Mathematical representations and expressions in students' text responses have been more difficult to identify and classify by the previous text analysis software packages the research team has evaluated (Goncher, Boles, Jayalath, 2016; Boles, Goncher, Jayalath, 2015). The clustering technique applied in this study can identify complete expressions, but requires more contextualisation for a meaningful evaluation of how the student referred to the expression in their explanation.

Word Embeddings

Word embeddings are simply real-valued vectors generally between 50 and 500 dimensions, depending on the training set. These vectors can be compared to each other by computing the cosine of the angle between them. Cosine similarity—and sometimes Euclidean similarity, which uses the length of the hypotenuse joining the two vectors—is the basis for establishing relationships between words. The method for training the words is parameterised, so that different features become more or less apparent. Representations for sentences are likewise real-valued vectors, often constructed in a compositional fashion from the vectors for words. The vectors are trained on some large body of text, for example Wikipedia. Using the word2vec tool (Mikolov, 2013), it is possible to train models on huge data sets.

The word2vec tool utilises the text corpus, i.e. the huge data set, as input in order to output the word vectors. Word2vec constructs a vocabulary from the input/ training text data, which it uses to produce the vector representation of words. The distance tool in Word2vec calculates the similarity between two words, or sentences, using the cosine similarity measure. Figure 1 provides a visualisation of the cosine distance model used to quantify word similarity for two words located in two documents.

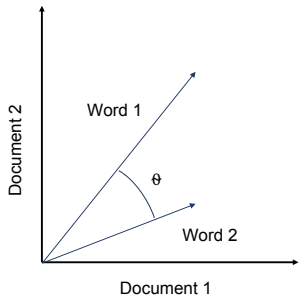

Figure 1: Word Cosine Similarity Model

As a numerical example, Table 3 shows most similar words to the given words “transform” and “shifted”. The distances are obtained using vectors trained on Wikipedia, and illustrates the potential outcomes of this approach, which analyses similarity through quantifiable measures. Vectors may be trained to enhance their associational qualities or their syntactic qualities. Training on different input text corpuses will also affect how relations between words are captured.

Table 3. Example Vector Similarities

Word	Cosine distance	Word	Cosine distance
transform	1	***shifted***	1
Transforms	0.7773872	Reverted	0.6257271
Transformation	0.63826245	Moved	0.6198429
transforming	0.63032836	Changed	0.61870444
fourier	0.61209697	Expanded	0.5978486
manipulate	0.611038	Transitioned	0.5875638
Transmute	0.6109377	Diminished	0.57048583
Mutate	0.5975203	Broadened	0.5685158
Transformed	0.5868472	Shifting	0.5508392
darkforce	0.57786304	Dwindled	0.54929495
shapeshift	0.5775432	waned	0.5470283

Multi-dimensional Scaling

One advantage of representing words and sentences as real-valued vectors is that it allows the application of other machine learning and visualization methods. Another method selected for analysing student responses is Multi Dimensional Scaling (MDS). This approach provides a means of clustering vectors in 2D space, allowing the evaluator to identify groups and outliers. Figure 2 illustrates how seven different example student responses from the Signals and Systems Concept Inventory textual data may be clustered based on the similarity of the words and sentences provided in each of their answers.

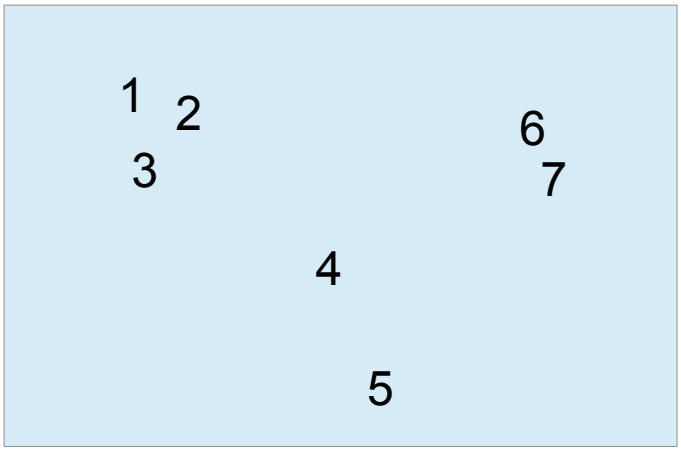

Figure 2: MDS representation of example student answers

From our analysis of example student responses illustrated in Figure 2, we see that student answers 1,2,3 are semantically more similar, and answers 1 and 5 would be considered less similar. We can use the clustering of vectors in 2D space to see the distances between student answers (new input) and compare how closely they are mapped to identified correct or more expert text responses to a given question.

Discussion

Based on our investigations of selected vector analysis approaches, we identified several difficulties, such as the initial time and resources required to develop and tailor the approach when automatically assessing student short answers, either for correctness or knowledge gaps. The main difficulties encountered when analyzing the students answers to a subset of the Signals and Systems Concept Inventory questions are: a) variation in vocabulary, b) variations in grammatical structures, c) precisely determining when specific concepts occur and don't occur, and d) taking into account relevant concept modifiers that may alter the assessment of the short-answer.

Proposed Software Techniques

A set of modules can be built to analyse student responses, and utilize command line programs that will process data in a specified format, such as comma separated value (CSV) files. More advanced applications based on these command line applications could be developed as part of future work.

Possible Modules

We suggest developing a set of modules to address the various features present in students' textual response that will provide information on the accuracy of students' answers. A list of module types is provided below:

- Abbreviations and Spelling Correction
- Recognize and normalize math symbols
- Thesaurus for different ways math entities may be expressed
- Quantifier detection, e.g. one, 1, single.
- Phrase splitting (Chunking)
- Guess, i.e. to detect guessing
- Summary statistics
- Concept thesaurus
- Paraphrase generator

- Answer clustering
- Semantic overlap of answers
- Retrieval of relevant online Wikipedia articles, or other relevant texts, for providing further context for questions and answers. Wikipedia can be used to train concept representations for judging the semantic overlap of responses.
- Simple visualizations of co-occurrence of concepts in responses

Conclusions

This study has focused on vector space representations for text, specifically in an engineering context. The suggested approach for an automatic evaluation system would make use of a combination of methods described in this paper, and can include classical Natural Language Processing (NLP) techniques.

The data analysed in the investigations presented in this paper consisted of student responses to short-answer questions in text form with associated metadata indicating the correctness for answers. We identified a number of difficulties faced when automatically assessing student short answers, either for correctness or knowledge gaps. One element, important for addressing these difficulties, is how words and sentences are represented in short-answer question responses. Semi-supervised approaches to analyze student textual data are still time consuming and require evaluation by an evaluator trained in, or familiar with, the text analysis methodologies. However, research needs to continue to address the need for developing efficient assessment techniques that can utilize advances in textual analysis software. Future work will address several overarching issues for an assessment system that utilizes students' written answers, such as addressing bias for students with strong written communication skills.

References

Aggarwal, V., S. Srikant, V. Shashidhar. (2013). Principles for using Machine Learning in the Assessment of Open Response Items: Programming Assessment as a Case Study, in: NIPS Workshop on Data Driven Education.

Bates, M. (1995). Models of natural language understanding. *Proceedings of the National Academy of Sciences of the United States of America*. *92 (22):* 9977–9982

Boles, W., Goncher, A., Jayalath, D. (2015). Uncovering Misconceptions through Text Analysis. *Proceedings of the 6th Research in Engineering Education Symposium*. Dublin, Ireland.

Goncher, A.M., Jayalath, D. and Boles, W. (2016) Insights Into Students' Conceptual Understanding Using Textual Analysis: A Case Study in Signal Processing. *IEEE Transactions on Education*, *59(3):* 216-223.

Goncher, A., Boles, W., & Jayalath, D. (2014). Using textual analysis with concept inventories to identify root causes of misconceptions. *2014 IEEE Frontiers in Education Conference Proceedings*. Madrid, Spain.

Jurafsky, D., and Martin, J.H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice Hall.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *ICLR 2013*.

Turney, P.D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research, (37):* 141-188.

Wage, K., Buck, J., Wright, C., and Welch, T. (2005). The Signals and Systems Concept Inventory. *IEEE Transactions on Education, 48(3):* 448-461.