Pointers to Conceptual Understanding

Samuel Cunningham-Nelson[a], Andrea Goncher[b], Michelle Mukherjee[a] and Wageeh Boles[a].
Queensland University of Technology[a], Charles Sturt University[b]
Corresponding Author Email: samuel.cunninghamnelson@qut.edu.au

CONTEXT

Concept inventories are tests used to elicit student misunderstandings and misconceptions. Traditionally, they exist as a set of multiple-choice questions (MCQs), including the correct option, as well as some distractors (Libarkin, 2008). This multiple-choice format allows for faster marking and feedback; however, it does not identify conceptual misunderstandings, or if a student has guessed the correct answer. By adding a space for students to add a textual justification (Goncher, Jayalath, & Boles, 2016), their answers can be checked to ensure that the concepts are correctly understood.

PURPOSE

Automated textual analysis will allow insights to be uncovered, and to help speed up the process of grading to give feedback to students and informing educators. As part of that process, we endeavour to address the following questions:

1. What pointers can be identified that indicate a student's conceptual understanding?
2. What conclusions can we make from these identified pointers to conceptual understanding?

APPROACH

Over the past four years, two concept inventories have been deployed, both with multiple choice questions, as well as a free text field for students to give reasoning and explanation. We will combine several machine learning techniques to analyse the textual response data, including:

- Word2vec – which allows words to be modelled as vectors, for easier computation (Mikolov, Chen, Corrado, & Dean, 2013)
- LDA (Latent Dirichlet Allocation) – Allows classification and grouping of topics and areas (Blei, et al., 2003)
- SVMs (Support vector machines) – which allow classification to be performed and similar areas grouped

RESULTS

Four pointers were identified to help to automatically determine if conceptual understanding is present. The first three pointers can be determined with certainty, the fourth "validity of the response" is one that is traditionally determined by a human marker. Comparing with an expert marked dataset, the algorithm to determine this pointer achieved a 75% accuracy.

CONCLUSIONS

Using the four identified pointers we are able to detect if a student has correctly identified the concept which they were being tested for in a particular question. The four pointers, allow some leniency if one of these is not achieved, and can also allow us to draw conclusions as to where issues lie in a student's understanding. This presents several opportunities for benefits such as individualised feedback for students and entire class feedback for educators.

KEYWORDS

Concept Inventories, Textual Analysis, Conceptual Understanding, Misconceptions, Machine Learning

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

Introduction

Formative assessment can be a strong contributor to enhancing students' learning outcomes, especially if these are used to provide them with meaningful and timely feedback. Nevertheless, for lecturers, this process can be very time consuming, and may even become impractical for large classes. One approach to reduce the marking load is to use Multiple Choice Questions, MCQs, which can be automatically marked. However, questions or assessments that require text-based answers can provide more information about students' understanding compared with standard MCQs (Birenbaum, 1987; Popping, 2012).

In order to better assess students' conceptual understanding, our study focuses on automating the collection and analysis of students' written textual responses, together with their MCQs selected answers. In our approach, we utilise text analysis and machine learning techniques to process the information gathered from students' textual responses.

Concepts

Concepts are representations of ideas in a simple form (Zirbel, 2006), and being the foundation or building blocks for an entire subject, they lie at the core of developing student understanding. Examples of concepts within the STEM area include: Time, Magnetism and Energy. These concepts, represented in a simple form, can appear easy to grasp however, many students fail to develop accurate understandings at school and can become confused and disenfranchised when successive ideas are introduced at university. Educators need to identify student misconceptions as they arise so that they can address them in their teaching. Understanding of concepts also allows for a deeper knowledge gain, as opposed to a more surface based approach. Concepts can also be defined in many ways, and this is just one example.

Assessing Conceptual Understanding

Assessing the conceptual understanding of a student can be a time consuming task, and responses need to be interpreted by a marker who is knowledgeable in the content area. Concept inventories were designed to help alleviate this issue, using a series of multiple-choice questions. They are designed to include the correct option, as well several distractors (Libarkin, 2008) however, one of the drawbacks of concept inventories is that they need to be designed by experts (Arbogast, 2016), and usually are designed to test specific concepts within an identified domain, e.g. signal processing.

Building on previous work (Cunningham-Nelson, Goncher, & Boles, 2016) further textual data has been gathered from another cohort of electrical engineering undergraduate students. We have selected a single question from the Signals and Systems Concept Inventory to investigate further.

Signals and Systems Concept Inventory

The Signals and Systems Concept Inventory (SSCI) was developed to assess core concepts in undergraduate signals and systems courses. The continuous-time and discrete-time versions are validated 25-question multiple-choice exams, which assess certain signal processing concepts in the continuous- and discrete- time domains. Potential solutions for every question include distractors (incorrect selections) that assist in determining the type of misconception a student may hold for each concept (Wage, Buck, Wright, & Welch, 2005). Developers of the SSCI determined and refined the distractor selections through the administration of earlier versions of the test. The SSCI was also designed to include a set of synthesis questions, which linked and built on several questions in the SSCI, and questions that require reverse reasoning. Additional details regarding the SSCI and its developers can be found at: http://signals-and-systems.org.

We utilised the discrete-time version of the SSCI in this paper, and examine how to accurately evaluate student conceptual understanding using the SSCI questions. In this study, students provided a multiple-choice response for each question and a written explanation as to why they selected the specific multiple-choice option. Previous studies utilising the augmented SSCI (multiple choice selection plus written response) have investigated the text evaluation processes and insights into students conceptual understanding not possible with MCQ-only questions (Goncher, Jayalath, & Boles, 2016; Boles, Goncher, & Jayalath, 2015).

The SSCI Discrete-time test has seven conceptual areas, i.e. math, linearity time invariance, sampling, filtering, transforms (time / frequency), convolution, and transform properties. We present an example from the SSCI Discrete-time to highlight example concepts, distractors, and student responses. Question 1 evaluates whether students can identify the sinusoid $\cos(\pi n)$ as having the highest frequency. Distractors include three signals that have obvious sinusoidal shapes, but tests if respondents confuse high amplitude with high frequency or large period, and if the sampling rate impacts the respondent's selection.

Question 1: The plots show segments of four periodic signals, all on the same time and amplitude scale. Each of the signals has the form $\mathrm{A}\cos(\omega_0 n)$ *with* $-\pi < \omega_0 \leq \pi$. Which signal has the highest frequency?

Question 1 on the discrete-time version is more difficult than the continuous-time version however 89% of respondents answered correctly. Correct example responses included, *"It has the shortest period and thus the highest frequency"* and *"The frequency is how fast something takes to complete one wavelength. A) takes 10s. C) takes 10s D) takes 20s B) takes <2.5s"*.

The example text responses illustrate how students can arrive at the correct multiple-choice answer, but have varying explanations. The first response highlights the relationship between frequency and period using the terminology, and the second example looks at each selection as a case of how long it takes before the signal repeats itself. One of the incorrect responses, e.g. *"amplitude of the cos wave is the frequency, the largest amplitude is the largest frequency,"* confirms potential misconceptions identified by the SSCI developers. Another respondent with an incorrect response, "*Has the highest and lowest points,*" also had the same misconception but did not use the specific terminology of amplitude in the text. The multiple-choice selection plus text responses show that students can arrive at either correct or incorrect answers, but may have varying ways of explaining the understanding, or misunderstanding of a concept.

Machine Learning and Text Analysis

In this work several text analysis and natural language processing techniques are used. These are combined with machine learning algorithms, to predict a particular outcome. Some key terms and processes used are discussed below.

LDA (Latent Dirichlet Allocation)

Latent Dirichlet Allocation is a probabilistic model for a collection of data, such as text (Blei, et al., 2003). Using a Bayesian technique data can be modelled and grouped. In terms of a text corpora, this means grouping topics and words together to obtain keywords. One common application for this, is automatically assigning labels to a large document which would otherwise need to be manually labelled.

Word2vec

Word embedding allows words to be modelled in a vector space. When viewing words in a vector space similar words will appear close to another, and relationships can be represented by addition and subtraction operations. To create the word vectors, a pre-existing model trained using many news articles was used (Mikolov, Chen, Corrado, & Dean, 2013). This model allows relationships between many English words to be preformed, and can be used to help with meaning in this analysis.

Research Questions

Taking advantage of the careful design of the concept inventory questions multiple choice questions and students’ free text justifications, we aim to identify key parts of responses, and checkpoints which might tell an educator about a student’s learning progress. Automated textual analysis will help speed up the process of grading and giving feedback to students and educators. As part of that process, we endeavour to determine:

1. What pointers can be identified that indicate a student’s conceptual understanding?
2. What conclusions can we make from these identified pointers to conceptual understanding?

Method

Pointers to Conceptual Understanding

Identifying students’ conceptual understanding is not a straight forward process. When a student’s free text response to a question is being marked, generally, a marker will have key aspects and terms in mind, and several “model solutions”. This however becomes more difficult when marking responses automatically. Having both the multiple choice and short response data available allows further insights into a student’s understanding. We have defined four pointers or indicators, which we believe, used together, will provide an indication of conceptual understanding.

Pointer 1 – Multiple Choice Correct

Utilising concept inventories which have carefully crafted questions and answer options allows common misconceptions to be identified using the multiple choice option selected by students. The multiple-choice response chosen is one pointer towards a students’ correct understanding of a concept. This multiple-choice option can be easily marked by a computer, and makes this first pointer straightforward to obtain.

Pointer 2 – Concept Mentioned

For this we need to first know which are the key concepts within the questions. This can be done either by defining these manually for the questions, or by using methods such as LDA to perform entity extraction automatically. After the topic is identified, we can then perform a keyword match to find the keywords that occur in particular responses. If the concepts are mentioned, then we can say this pointer has been met.

Pointer 3 – Response Uncertainty

In their responses, students were asked to use words such as ‘guess’ or process of ‘elimination’ to explain how they come to their answer. If these words are mentioned within a students’ response, this adds doubt to the level of certainty and confidence in their answer. This is something that is important to consider when performing analysis on a student response. It is also important to consider the difference between the two words. We considered that responses which include the word “guess” are more uncertain than those which have the word “elimination”.

Pointer 4 – Free Text Validity

This is the most difficult pointer to determine, and results will vary. The validity of the written response would traditionally need to be evaluated by a human marker. The marker will compare the given response to a model or bank of model responses or their down expert knowledge of the subject. However, this needs to be done in an automated fashion. Using several machine learning methods, we aim to replicate this manual process.

Responses for Q1 of the SSCI were initially manually labelled (marked) into three categories:

1. Concepts mentioned and correctly used
2. Concepts mentioned, but incorrectly used, or incorrect
3. Answer incorrect or major misconception.

These responses were manually labelled to be used to train and validate the machine learning algorithms evaluated for this task. We start with the text responses given by students and perform some initial pre-processing of the text to ensure that the text is ready for analysis. This involves: transforming all the text to lowercase, moving stop words (i.e. it, and, the) and lemmatising the words (using the base of each word).

The sentences are then converted into a “bag of words” model for processing. A bag of words model (word frequency model) means that each sentence is converted into a row of ones and zeros. All the sentences together form a sparse matrix, which can be quite large however this sparse matrix is the input into various machine learning classification methods. This bag of words representation is then passed into various machine learning classifications methods mentioned below (Boser, Guyon, & Vapnik, 1992; Ray, 2017). The preliminary results can be seen in Table 1.

Table 1 – Bag of Words Accuracies

Method	Extra Trees Classifier	Linear Discriminant Analysis	Logistic Regression	KNN	DT	Naïve Bayes	Linear SVM	Gaussian SVM
Accuracy	68.9%	58.0%	71.8%	64.9%	64.9%	56.9%	71.2%	**74.1%**

Table 1 shows that the Gaussian SVM classifier produces the best results (most correctly classified responses). Using word2vec, representing the word responses as vectors, the average word vector for each response can be found. This word vector was used with the Gaussian SVM classification method above. This achieved a classification accuracy of **77.0%**. This classification model was then used for predicting the outcome autonomously for pointer 4.

Results

Overall Results

Initially results were examined for each pointer separately. These results have been summarised in four separate tables, to reveal how the group of students performed across each pointer for Q1 of the SSCI.

Pointer 1 – Multiple Choice Correct

The multiple-choice results show a good initial result for class understanding as a whole. Table 2 below shows a count and percentage for both incorrect and correct results from the multiple-choice answers. We can see that most students answered this question correctly, and would hope that these students understand the concepts in the question. We can investigate this further, looking at the remaining three pointers.

Table 2 - Summary of Multiple Choice Results (N=174)

	Count	**Percentage**
Correct	155	89.1%
Incorrect	19	10.9%

Pointer 2 – Concept Mentioned

The second pointer identified was whether students mentioned the key concepts for the particular question. The key concepts that were required in a response could be determined in one of two ways. One option is for the concepts to be identified by someone familiar with the topic area or subject. For example, for the chosen question, three important concepts were identified: Frequency (freq), Period and Time.

The second option is to use the LDA method to automatically identify topics within a document can be identified. Providing all the student responses as input into the LDA algorithm, topics can quickly be extracted. In this case, the top three topics grouped by single words identified were: “period”, “signal” and “b”. Interestingly if the topics are grouped into a larger number of words, further patterns emerge, such as the words “highest, frequency, and b” being grouped into one topic. These show key terms which we might expect in a correct response.

Table 3 shows a count of the responses which mentioned the manually chosen keywords: frequency and period. Interestingly this number is significantly smaller than the number of students who got the multiple choice option correct. Whether the student mentioned one of the desired concepts is another pointer for correct understanding.

Table 3 - Summary of Concept Mentioned Results (N=174)

	Count	Percentage
Concept Mentioned	123	70.7%
Concept not Mentioned	51	29.3%

Pointer 3 – Response Uncertainty

A further pointer for students’ conceptual understanding is the certainty in their answer. If a response mentions “guess” or “elimination”, it can indicate little or no confidence in the response. Doubt expressed in a student response could indicate a possible misconception, or a lack of complete conceptual understanding. Table 4 shows a summary of these results, with a breakdown of certainty within the various levels. It can be seen from this table that most students are certain in the answer that they select.

Table 4 - Summary of Uncertainty Results (N=174)

		Count		Percentage	
Students Uncertain	**Elimi**	1	7	0.6%	4.0%
	Guess	6		3.4%	
Students Confident		167		96.0%	

Pointer 4 – Free Text Validity

The final pointer towards assessing conceptual understanding, and arguably the most important is the free text response written by the student. A valid response from a student is one that demonstrates full conceptual understanding, whereas misconceptions or a lack of understanding can also be determined. Using the prediction methods previously discussed, Table 5 shows a summary of these results. These provide an overall picture of the understanding of the desired concepts in this question.

Table 5 - Summary of Validity Results (N=174)

	Count	Percentage
Concept correctly used	130	74.7%
Concepts mentioned, but incorrectly used	20	11.5%
Answer incorrect or major misconception.	24	13.8%

Selected Examples

Overall analysis can provide a good pointer to the overall level of understanding in a group of students and looking at individual responses allows conceptual understanding to be examined on a student by student basis. Selected examples of responses have been chosen to show the possible conclusions that can be drawn from the analysis. For each example, the automated outputs will be given, and explored. These chosen examples will hopefully demonstrate where the automated process can succeed, but also where it can be improved.

Selected Example 1

The first example selected, given by a student is, *"guess"*.

Using the automated methods above, the following outcomes are achieved for the four identified criteria as shown in Table 6.

Table 6 - Selected Example 1 Automated Results

P1: Multiple Choice	P2: Concept Mentioned	P3: Certainty	P4: Explanation "valid"
✗	✗	✗	✗

From these four criteria, we can reasonably determine that the student had no understanding of the required concepts. This is evident by their short response, and no explanation.

Selected Example 2

The second response selected is, *"Frequency is defined as number of cycles per second. plot b has the most number of cycle within a one time period."*.

Table 7 shows the outcomes for each of the four pointers. Each of the pointers to conceptual understanding has been met, demonstrating that the student understands the concept being tested. This can be verified by reading the students response and comparing it to the previously given model answer.

Table 7 - Selected Example 2 Automated Results

P1: Multiple Choice	P2: Concept Mentioned	P3: Certainty	P4: Explanation "valid"
✓	✓	✓	✓

Selected Example 3

The third student response selected is, *"Most changes between pos & neg in given time scale"*.

The results in Table 8 from the four pointers show that the student correctly met the first three pointers, but did not meet the final one. However, upon reading their response, we can say that their explanation is valid and demonstrates understanding even though this response is quite different from the "typical response" that is expected. Therefore, a response obtaining the first three pointers, but missing the third should be manually reviewed to check the automated classification of the third response.

Table 8 - Selected Example 3 Automated Results

P1: Multiple Choice	P2: Concept Mentioned	P3: Certainty	P4: Explanation "valid"
✓	✓	✓	✗

Selected Example 4

The final response selected is, "It has the highest density of wave"

Table 9 shows that the student met two out of the four pointers outlined for conceptual understanding. They did select the correct multiple choice option, and expressed no doubt about their answer, however they did not mention any of the listed concepts, and their explanation was not deemed to be correct. This indicates a possible need to reinforce required concepts.

Table 9 - Selected Example 4 Automated Results

P1: Multiple Choice	C2: Concept Mentioned	C3: Certainty	C3: Explanation “valid”
✓	✗	✓	✗

Conclusions From Combinations of Pointers

The four selected examples and out conclusions are summarised in Table 10. When trying to determine conceptual understanding, the information from each of the four pointers can be used. A few combinations have just been chosen to demonstrate the four pointers listed here.

Table 10 - Combinations and Conclusions from Pointers

P1: Multiple Choice	P2: Concept Mentioned	P3: Certainty	P4: Explanation “valid”	Overall Conclusion
✗	✗	✗	✗	No understanding at all of concept
✓	✗	✓	✗	Possible misconception, since they have keywords or a correct response
✓	✓	✓	✗	The first three pointers lead to a need to double check the text response manually
✓	✓	✓	✓	Student has full understanding of concept

Conclusions and Recommendations

This paper presents four pointers identified to assess conceptual understanding. Data was gathered across a four-year period, using a multiple choice concept inventory with added text responses. Using the four pointers identified we can make conclusions about the understanding of the student for the particular question. All of the four pointers are automatically evaluated using a combination of text analysis techniques and machine learning methods. The first three points can be determined with certainty, the fourth “validity of the response” is one that is traditionally determined by a human marker. Compared with an expert marked dataset, the algorithm to determine this pointer achieved a 75% accuracy. One interesting note to make, is that the number of students who selected the correct MC option is significantly more than the number of students who explained in words the correct response. This emphasises that the combination of MCQs and short responses helps to test conceptual understanding.

We have conducted our investigations on one question as an initial study. Further work includes looking at how other types of models may help to improve the prediction accuracy for pointer 4. Models such as recurrent neural networks take word order into account, which our current prediction model does not. It would also be beneficial to consider ways which the combinations of pointers present or not present can be used to give individual feedback to students.

Using text analysis and machine learning methods, we were able to assess to a certain degree, a student’s conceptual understanding of the presented topic. Using the four identified pointers we are able to detect if a student has correctly identified the concept they were being tested for in a particular question. This presents several opportunities for benefits such as individualised feedback for students and entire class feedback for educators.

References

Arbogast, C. (2016). Assessing Student Conceptual Understanding: Supplementing Deductive Coding with Natural Language Processing Techniques.

Blei, D. M., Edu, B., Ng, A. Y., Edu, A., Jordan, M. I., & Edu, J. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 993-1022.

Boles, W., Goncher, A., & Jayalath, D. (2015). Uncovering Misconceptions through Text Analysis. 6th Research in Engineering Education Symposium (REES2015). Dublin, Ireland.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152). ACM.

Buck, J. R., & Wage, K. E. (2005). Active and Cooperative Learning in Signal Processing Courses. IEEE Signal Processing Magazine, 22(2), 76-81.

Cunningham-Nelson, S., Goncher, A., & Boles, W. (2016). A three-year longitudinal textual analysis investigation of students' conceptual understanding: Lessons learnt and implications for teaching. AAEE2016. Coffs Harbour, NSW.

Goncher, A. M., Jayalath, D., & Boles, W. (2016). Insights Into Students' Conceptual Understanding Using Textual Analysis: A Case Study in Signal Processing. IEEE.

Libarkin, J. (2008). Concept inventories in higher education science. BOSE Conf.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Nips, 1-9.

Ray, S. (2017, September 9). Analytics Vidhya. Retrieved from Essentials of Machine Learning Algorithms: https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

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

Zirbel, E. L. (2006). Teaching to promote deep understanding and instigate conceptual change. Bulletin of the American Astronomical Society, 1220.

Acknowledgements

We would firstly like to acknowledge the authors of the Signals and Systems Concept Inventory for allow us to use it in this study, and the students who provided responses to the questions. The authors would also like to thank Dr Dhammika Jayalath for his efforts in implementing this concept inventory over several years.