ATLANTIS PRESS
Advances in Social Science, Education and Humanities Research (ASSEHR), volume 68
2017 International Conference on Advanced Technologies Enhancing Education (ICAT2E 2017)

Alignment of Standards using WordNet for Assessing K-12 Engineering Practices in a Participatory Learning Environment

Sam Shuster[1, a], Erin Shaw[1, b]

[1]University of Southern California, Los Angeles, CA, USA

[a]sshuster@edmunds.com, [b]erinshaw@usc.edu

Keywords: K-12 educational assessment, Next Generation Science Standards, Science and Engineering practices, participatory learning, machine learning, latent semantic analysis, WordNet.

Abstract. Student assessment based on advanced computational methods will play a critical role in 21st century educational practices. In this paper we describe the challenge of automating the instructional assessment of student discourse based on national standards, in the context of an experimental participatory learning platform. Latent Semantic Analysis, machine learning, data mining and natural language processing techniques were used in conjunction with WordNet to create a classification scheme for engineering standards based on the *Science and Engineering Practices* in the U.S. Next Generation Science Standards. The scheme was applied to interactive student presentations, to assess and report on students' engineering and domain learning.

Introduction

Participatory cultures are defined as cultures with relatively low barriers to artistic expression and civic engagement, strong support for creating and sharing one's creations, and some type of informal mentorship [1]. Participatory cultures foster new media literacies that build on traditional literacy skills taught in the classroom but focus on social skills developed through collaboration and networking, such as play, performance, simulation, appropriation, distributed cognition, and judgment. The affordances of Web 2.0, associated with social software such as blogs, folksonomies, and peer-to-peer media sharing, and the ubiquity of networked computers in K-12 schools have made possible new educational practices that have the potential to produce "radical and transformational shifts" in learning [2]. However, Pedagogy 2.0 has been realized only in fits and starts due to inherent technological and educational challenges.

For new educational practices to become accepted, they must fit into the curriculum, be aligned to state standards and have appropriate assessments [3]. Reilly et al. [4] have focused on how contemporary educational practices embrace participatory learning by developing an online learning platform that implements the four C's of participation in the learning process: create, circulate, collaborate and connect. Their resulting *Playground*, an implementation of the *PLAY! –Participatory Learning and You!* framework, is a social, multimedia development environment that encourages users to experiment with ideas. It has been used for teacher professional development and is well positioned for instructional classroom use. The challenge now is how to evaluate the learning effectiveness of using Playground.

This research focuses on integrating standards-aligned engineering education assessments into Playground. The work will be of interest to those wishing to implement participatory learning practices in an education setting. We review our initial effort to implement as assessment tool, identify the challenges encountered, and describe an exploration of computational methods to address one particular challenge, the alignment of standards. Latent Semantic Analysis (LSA) and WordNet were used to develop a systematic method for classifying the extent to which engineering standards based on the *Science and Engineering Practices (SEP)* in the Next Generation Science Standards (NGSS) [5] appear in student text.

Data and Methodological Approach

Preliminary data consisted of Playground text from six students between the ages of 10 and 20 whose discussions focused on the domain of Minecraft [5]. The engineering standards used were grouped based on the NGSS *SEPs* [6]. The eight final categories were "Analyzing and interpreting data", "Asking questions and defining problems", "Constructing explanations and designing solutions", "Developing and using models", "Engaging in argument from evidence", "Obtaining, evaluating, and communicating information", "Planning and carrying out investigations" and "Using mathematics and computational thinking". Scikit-learn [7] and NLTK (Natural Language Toolkit) [8] were used for the machine learning and natural language processing algorithms.

Table 1. Base processing steps to create a vector of tokens.

Step	Example
1. Raw Text	*Planning and conducting collaboratively to produce*
2. Tokenization & Conversion to Lower Case	['planning', 'and', 'conducting', 'collaboratively', 'to', 'produce']
3. Lemmatization	['plan', 'and', 'conduct', 'collaboratively', 'to', 'produce']
4. Stop Word Removal	[plan', 'conduct', 'collaboratively', 'produce']
5. NGram Generation (Bigram here)	['plan', 'conduct', 'collaboratively', 'produce', 'plan conduct', 'conduct collaboratively', 'collaboratively produce']

Raw text must undergo a series of processing steps to render it to a quantitative, structured format that can be analyzed automatically. The processing of text is highly application specific and can have huge implications as to the quality and validity of the results. The base processing steps are outlined in Table 1. The result is a list of plain text tokens of which a number of analysis techniques can be applied to determine similarity between documents. One such technique is Latent Semantic Analysis (LSA) [9, 10]. LSA is a non-supervised technique that relies on three steps to cluster documents based off of hidden semantic similarity. The first step is to apply a transformation to the list of text tokens to generate feature vectors of numbers; the second step is to reduce the dimensionality of the features; and the last step is to cluster the data using some clustering algorithm or to compute the cosine distance. Common choices for the above three steps would be to use Term Frequency multiplied by Inverse Document Frequency (TFIDF) to generate the feature vectors, Principal Component Analysis (PCA) to reduce the dimensionality of the feature vector and K-Means clustering to cluster the documents. Using this methodology, the goal will be to determine the similarity of a student's textual content to each of the eight standards by analyzing each student individually in conjunction with the complete set of standards at a time. As an example, let's examine three documents of preprocessed text:

- Document 1 (D1): ['plan', 'conduct', 'phenomenon']
- Document 2 (D2): ['describe', 'phenomenon']
- Document 3 (D3): ['conduct', 'relationship', 'cooperation']

Table 2 shows the resulting term frequency feature vectors of the three sample documents listed above. See that the resulting matrix will be an N by M matrix where N is the number of documents and M is the number of unique terms. The TFIDF transformation is utilized to offset the importance of common words by normalizing the counts based off of how many unique documents the word appears in. In this example, we can observe that this process provides a way to differentiate terms more readily then simple term counting.

Table 3A shows the resulting matrix after PCA was used to reduce the dimensionality of the feature set to two. At this point the dimensions do not have any specific definition anymore and instead correspond to a linear mapping that represents the maximum amount of variance in the original data. Finally, in Table 3B, the cosine similarity is calculated of the PCA feature vectors. The idea is that given two feature vectors of length D that describe two documents, one can calculate the cosine of the angle between the two vectors in D-dimensional space. A cosine of zero means that the two feature vectors are orthogonal and that the two documents are completely unrelated. A cosine of

one means that the two vectors are parallel and that they are identical. In this example, D2 and D3 are found to be completely unrelated, while D1 is found to be more similar to D2 then to D3. While this makes some sense, we would still expect that if D1 is similar to D2 and D1 is similar to D3 that there must exist some similarity between D2 and D3. Thus, the LSA analysis very much depends on the vocabulary utilized in the documents regardless of semantic content.

Table 2. TDIF feature vectors.

	conduct	cooperation	describe	phenomenon	plan	relationship
D1	0.52	0	0	0.52	0.68	0
D2	0	0	0.80	0.61	0	0
D3	0.47	0.62	0	0	0	0.62

Table 3A,B. PCA reduced feature vectors and their cosine similarities.

	D1	D2
D1	0.09	0.64
D2	0.66	-0.38
D3	-0.75	-0.26

	D1	D2	D3
D1	1	0.31	0.25
D2	0.31	1	0
D3	0.25	0	1

Using WordNet to Align Standards

As mentioned earlier, at the crux of LSA analysis is the comparison of term occurrences between documents. This analysis is highly vulnerable to synonymy, meaning that the presence of synonyms in two different documents will not improve their similarity scores. To demonstrate that this is obviously a prevalent problem when attempting to align the standards with students text, the Flesch-Kincaid readability score [11], which is a standard metric for measuring the expected grade level of a writing sample, was applied to all students text and standards text to provide some inference for vocabulary level and is shown in Table 4.

Table 4. Flesch-Kincaid readability scores comparisons: User conversation versus standards.

Student Participant	P1	P2	P3	P4	P5	P6	Average
Flesch-Kincaid Readability	2.1	2.4	8.4	7.7	11.3	6.01	6.32

Standard	S1	S2	S3	S4	S5	S6	S7	S8	Average
Flesch-Kincaid Readability	13.8	5.5	15.3	9.6	16.6	17.8	16.8	14.1	14.94

The average grade level of the standards text is as expected almost twice that of the students. From this we can infer that once stop words have been removed the vocabulary overlap between the two sources will be quite small, which poses a challenge to our semantic similarity tools. One way to assuage this issue is to attempt to map words from both the student and standards domain to a common vocabulary. The challenge will be to conduct the mappings in such a way that preserves the originally intended semantic connotations. This can be tricky when dealing with words such as ‘make’ that have many different meanings, which is an inherent difficulty with this approach.

A database called ‘WordNet’ [12, 13] has the lexicon and semantic relationships encoded that is most apt for a task such as this. At the core of WordNet are atomic blocks called lemmas. Lemmas are of the form ‘word.POS.sensenum.sensestring’ where *POS* is the part of speech, *sensenum* is a number that disambiguates between words that have different meanings for the same part of speech, and *sensestring* is the actual string that corresponds to this specific meaning [12]. Synsets can be thought of as synonyms. WordNet provides functional tools that operate on lemmas and synsets to discover such useful characteristics as antonyms, troponyms and entailment but depends on the part of speech of the word. So ideally, if the correct semantic synset could be found for a word, that word or subsets of words could be included or substituted in its feature vector. Unfortunately, the way the standards are written makes the accuracy of the automatic POS tagging quite low and, because WordNet is does

not have the completeness of an English dictionary, there will be some cases where no can be lemmas are found. Let us look at a sample list of tokens from the standards and their corresponding synsets:

- ***plan*** *[Synset('plan.n.01'), Synset('design.n.02'), Synset('plan.n.03')]*
- ***conduct*** *[Synset('behavior.n.01'), Synset('demeanor.n.01')]*
- ***investigation*** *[Synset('probe.n.01'), Synset('investigation.n.02')]*
- ***collaboratively*** *[]*
- ***produce*** *[Synset('produce.v.01'), Synset('produce.v.02'), Synset('produce.v.03'), Synset('produce.v.04'), Synset('grow.v.07'), Synset('produce.v.06'), Synset('grow.v.08')]*

The first three words were incorrectly tagged as nouns, which resulted in synsets such as 'demeanor' for the word 'conduct' that are incorrect. Furthermore, the word *collaboratively* does not appear in the WordNet dictionary, yet 'collaborate' does. Thus, in order to maximize the number of words from the standards that are correctly mapped to a semantically correct synonym set, some additional preprocessing techniques had to be used to a) identify the correct POS tag and b) morphologically stem words that cannot be found in WordNet. To deal with the POS tagging issue, as educational standards can be thought of as a list of requirements that detail what a student should be able to do, a basic heuristic could be to think of the standards as in essence a group of verbs. Therefore, the strategy used was to attempt to convert all non-stop words to verbs. To deal with the problem of not finding a word in WordNet, the most similar lemma out of all lemmas in WordNet was substituted for the original word. Similarity between words was measured using the edit distance and an arbitrary threshold was set to determine when a word was similar enough to substitute it for the original.

LSA was applied to six sets of feature matrices where each feature matrix consists of 9 feature vectors (1 student with 8 standards) that were generated using no WordNet synonyms, and with WordNet synonyms added. These results are summarized in Figure 1. Each spoke of the radar chart corresponds to a specific standard (S1-8), and each axial line corresponds to the cosine similarity score (-1.0 to 1.0) of that student participant (P1-6) with that standard. With the absence of human annotated ground-truth values, evaluating the validity of the measurements is difficult, however, we can comment on the effect that the inclusion of the WordNet synonyms has. Firstly, observe that the Cosine similarity scores without the use of WordNet seem to be able to only distinguish two groups of users (P1, P2, P6 versus the rest). This in contrast with the results with WordNet synonyms, which is able to give each of the students a separate diagnostic. These findings support the notion that the transformed vocabulary of the two domains has a much higher level of overlap. Of great interest would be to recalculate the Flesch-Kincaid readability scores of the transformed text. Unfortunately, this would not be valid due to the dependence of that metric on sentence length, which would be dramatically changed with the inclusion of synonym sets.

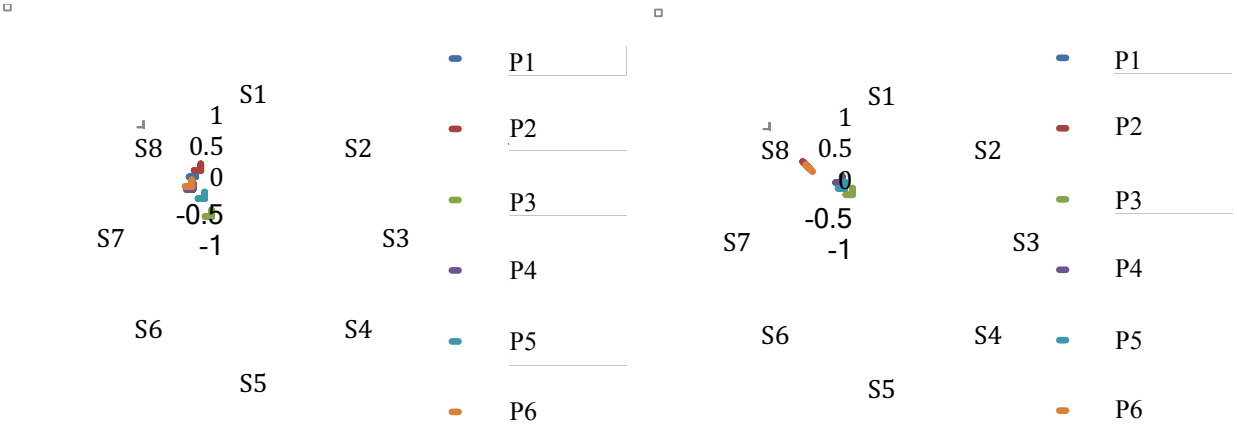

Figure 1. Cosine similarity scores with WordNet (left) and without WordNet (right) synonyms

Summary

Research on advanced computational methods will play a critical role in the assessment of 21st century practices. In this paper, we closely analyzed how new science and engineering content standards might be used to assess participatory student practices in education. Initial results of aligning standards to student text through the use of LSA coupled with WordNet are promising. Further exploration will require annotated ground-truth and more data with larger focus groups. We also continue to explore the student analytics pipeline with respect to new engineering standards.

Acknowledgments

The authors thank Creative Director Erin Reilly and Software Developer Aninoy Mahapatra of USC's Game Innovation Lab for their support and assistance. This material is based upon work supported by the National Science Foundation under grant #1008747, and a grant from USC's Undergraduate Research Associates Program (URAP). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Literature References

[1] Jenkins, H., Clinton K., Purushotma, R., Robinson, A. J., & Weigel, M. (2006). Confronting the challenges of participatory culture: Media education for the 21st century. Chicago, IL. Online at http://www.macfound.org/media/article_pdfs/JENKINS_WHITE_PAPER.PDF.

[2] McLoughlin, C., & Lee, M. J. (2007). Social software and participatory learning: Pedagogical choices with technology affordances in the Web 2.0 era. In *ICT: Providing choices for learners and learning. Proceedings ASCILITE, Singapore 2007*, pp. 664-675.

[3] Nobori, M. (2013). A Step-by-Step Guide to Best Projects: Discover a project-based learning model that motivates students to pursue knowledge and drives academic achievement, Edutopia. Online at http://www.edutopia.org/stw-project-based-learning-best-practices-guide.

[4] Reilly, E., Jenkins, H., Felt, L.J., & Vartabedian, V. (2012) *Shall we PLAY?* Online at http://www.annenberglab.com/sites/default/files/uploads/Shall_We_PLAY_final_small.pdf.

[5] Shaw, E., La, M. Phillips, R. & Reilly, E. (2014) PLAY Minecraft! Assessing secondary engineering education using game challenges within a participatory learning environment. In *Proceedings, American Society for Engineering Education (ASEE) 2014*.

[6] Next Generation Science Standards (2013). Online at http://www.nextgenscience.org.

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B, Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research vol. 12, pp. 2825-2830.

[8] Bird, S., Loper, E., & Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc.

[9] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. JASIS, 41(6), 391-407.

[10] Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2013). Handbook of latent semantic analysis. Psychology Press.

[11] Kincaid J.P., Braby, R., Wulfeck, W.H. II (1983). "Computer aids for editing tests". *Educational Technology*. **23**: 29–33.

[12] Christiane Fellbaum (1998) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[13] WordNet (2010). Princeton University. Online at http://wordnet.princeton.edu 3