Modeling Classifiers for Virtual Internships Without Participant Data

Dipesh Gautam
The University of Memphis
Memphis, TN 38152
dgautam@memphis.edu

Zachari Swiecki, David W. Shaffer
University of Wisconsin-Madison
Madison, WI 53706
{swiecki,dws}@wisc.edu

Arthur C. Graesser, Vasile Rus
The University of Memphis
Memphis, TN 38152
{graesser,vrus}@memphis.edu

ABSTRACT

Virtual internships are online simulations of professional practice where students play the role of interns at a fictional company. During virtual internships, participants complete activities and then submit write-ups in the form of short answers, digital notebook entries. Prior work used classifiers trained on participant data to automatically assess notebook entries from these learning environments. However, when teachers create new internships using available authoring tools, no such data exists. We evaluate a method for generating classifiers using specifications provided by teachers during their authoring process instead of participant data. Our models rely on Latent Semantic Analysis based and Neural Network based semantic similarity approaches in which notebook entries are compared to ideal, expert generated responses. We also investigated a Regular Expression based model. The experiments on the proposed models on unseen data showed high precision and recall values for some classifiers using a similarity based approach. Regular Expression based classifiers performed better where the other two approaches did not, suggesting that these approaches may complement one another in future work.

Keywords

Automated assessment, text classification, LSA, neural network, semantic similarity, regular expressions,

1. INTRODUCTION

Recently, authoring tools have been developed that let teachers customize and create new versions of digital learning environments such as intelligent tutoring systems and simulations [15]. However, if these environments use integrated automated systems, such as classifiers, customization can be problematic: a new environment invalidates previous automated systems and participant data does not yet exist to train new ones. Therefore, teachers who author these learning environments must implement them, at least initially, without a key component of the technology.

For example, virtual internships are online simulations of professional practice where participants play the role of interns at a fictional company [14]. During virtual internships, participants complete activities and submit work in the form of digital notebook entries. Typically, these are short answer responses ranging from a few sentences to a paragraph in length. Prior work has investigated automated assessment of notebook entries by training classifiers on participant data [10]. However, since the development of the Virtual Internship Authoring Tool [18], teachers can now customize activities and their notebook requirements. Thus, previously developed classifiers may no longer be valid and, initially, participant data is not available to use for model training.

In this paper, we present and test a method that addresses this issue by generating classifiers from specifications that teachers provide during the authoring process rather than waiting to generate them from participant data. Ultimately, these classifiers will be integrated into a fully automated assessment system that will score participant notebook entries. In this study, however, we only report on the development of classifiers for determining whether teacher defined requirements are present or absent in an entry, not classifiers that assign a final assessment.

2. BACKGROUND

Several automated essay scoring systems [3, 8, 16] have been developed to tackle the challenges of costs, reliability, generality and scalability while assessing open-ended essays. Previous researches on automated essay scoring focused on the argumentative power of an entire essay, while in our case, the student generated content is typically short text the length of a sentence or paragraph. Also, the focus of our assessment is to classify the content based on the presence or absence of semantic content defined by teachers during their authoring process. This means that style and higher-level constructs, such as rhetorical structure, are less important in our task compared to essay scoring and that factors that focus more on content measures are more important. Therefore, we limit our work to a semantic similarity approach and Regular Expression (RegEx) matching approach to identify the presence of targeted semantic content in participant generated text.

Various methods of text similarity measures have been used from the very early years of information retrieval. One of the simplest approach is to use the lexical overlap between the texts, however this approach does not consider the semantic relation between the words. Salton & Lesk [13] used is term frequency based vector model for documents similarity. Such model fails when two texts with same meaning have few overlapping words. Other approaches use knowledge base such as WordNet to find semantically similar words in two text [4, 9]. However these approaches face challenges of word sense disambiguation. Other approaches use LSA or LDA methods that rely on large corpus and do not face word sense disambiguation challenge [11].

Rus et al.[11] collected a large corpus of student-generated paraphrases and analyzed them along several dozen linguistic dimensions ranging from cohesion to lexical diversity obtained from Coh-Metrix [5]. They used the most significant indices to build a prediction model that can identify true and false paraphrases and also several categories of paraphrase types. Our work is significantly different than their work as our classifier

model does not rely on participant generated content (we develop classifiers from teachers specifications of content before any participant response is available), secondly our paraphrase detection model measures semantic relation between the text without depending on linguistic features such as content word counts.

Our LSA based similarity method relies on the combination of constituent words a phrase. Hence the similarity score will be more biased towards phrases having common words. While the Neural Network (NN) based semantic similarity method proposed by [7, 17], which we also explored, projects the phrase pairs into common low dimensional space hence the similarity score obtained will be more consistent irrespective of the presence of common words in the phrases.

Our work closely relies on previous works [2, 4, 9] where the authors proposed methods to measure the semantic similarity between texts. The authors in [2] and [4] used knowledge bases such as WordNet while the authors in [9] used word to word similarity and vectorial representation of words derived using Latent Semantic Analysis (LSA) to compute the semantic similarity of two given texts. In addition to these methods, we used in our work presented here phrase vectors generated using Neural Network based models [7, 17]

Our work is also partially related to the work by Cai et al.[1], which proposed methods to evaluate student answer in an intelligent tutoring system. They used LSA and RegEx to assess student answers. Their work showed that the carefully created RegEx had high correlation with human raters' scores. They also noted that the correlation increased when the expected answers created by experts were combined with the previous students' answers to assess new student answers.

3. METHODS

We developed three different types of classifier models and evaluated their performances separately.

To generate our classifiers, we worked with data from one teacher as she authored an activity in the virtual internship, *Land Science*. In *Land Science*, participants work to design a city zoning plan that balances the demands of stakeholders who advocate for indicators of community health. In the activity that this teacher customized, participants describe their proposed zoning changes in a notebook entry. In the first step of our method, the teacher defines assessment criteria for an entry in terms of core concepts, or the key semantic content they want to be present or absent in an entry. For this entry, the teacher defined five core concepts (see Table 1). Next, she constructed six example entries and identified the chunks of text in each example that expressed each concept. In addition, she provided lists of keywords for each core concept that she expected to be present in participant notebook entries.

Afterward, we developed various classifiers for each core concept based on the teacher provided items: sample responses, core concepts, and concept keywords. In this paper, we report three such classifier types; The LSA based semantic similarity threshold classifier, the NN based semantic similarity classifier, and the RegEx based classifier.

In both the LSA based and NN based classifiers, we use a sliding window to search for the most similar chunk in an intern's notebook entry. That is, for each teacher-defined chunk, we slide a window of equal size over the student entry. For each such participant-chunk identified by the sliding window over the student's notebook entry, we calculate the semantic similarity of the text within the window to the teacher-defined chunk. After the similarity of all windows to a teacher-chunk has been calculated, we assign the highest value as the similarity score for a given core concept. For LSA based classifiers, we calculated the similarity score using SEMILAR [12]. For the NN based classifier, we calculated similarity score using the Sent2Vec[1] tool. Since both the tools are capable of taking phrases or sentences as input, we give the chunks as input phrase, hence in the rest of the sections, we call these chunks as phrases.

If the highest similarity score is high enough, e.g. higher than a threshold, we decide the target core concept is present in the student response. Otherwise, we infer the student respond does not include the core concept. That is, we developed a semantic similarity based classifier for assessing students' responses.

In order to choose a threshold for the similarity based classifiers, we derived a threshold by calculating the similarity score between the chunks of each of the core concepts tagged by the teacher for both LSA based and NN based methods. See the experiment section for details.

To test the validity of our approach, we developed classifiers for each target concept and then tested them using 199 participant entries coded by humans for the presence or absence of each core concept.

Because our initial thresholds were created without the aid of participant data, we expected that better thresholds would exist. We therefore sought to compare the performance of our classifiers using two different thresholds, the *derived* thresholds above and *ideal* thresholds (described in more detail below). To calculate the ideal threshold for each classifier we varied the semantic similarity thresholds from zero to one and obtained precision and recall measures for each threshold using participant data.

For the RegEx based classifiers, we used the teacher provided keywords, which were generated without using participant data, to create regular expression lists for each core concept. We infer that the target core concept is present in a given entry as long as any of its associated keywords are present, as determined by regular expression matching. Therefore, in contrast to the LSA and NN models, a threshold is not required for the RegEx classifiers.

The semantic similarity approach minimizes the teachers' input which encouraged us to adopt it for assessing participant responses with respect to containing (or not) targeted, required concepts. This method is also relatively easy to automate, meaning that after the teacher has made a small set of specifications, classifiers can be developed without further human input. The RegEx approach is less flexible compared to the semantic similarity approach as novel expressions of a core concept, not encoded yet in the regular expressions, are less likely to be correctly identified. However, the RegEx is capable of identifying core concepts that are characterized by a closed set of keywords and semantic similarity may not be able to perform as needed.

[1]https://www.microsoft.com/en-us/download/details.aspx?id=52365

4. EXPERIMENTS AND RESULTS

First, we describe the data set we used in our experiments and then present the results obtained with our automatically generated classifiers. We also apply these classifiers to participant generated notebook entries to assess the performance of our models on unseen data.

4.1 Data Set

As we mentioned above, our classifiers were generated from specifications made by a teacher as she customized an activity in *Land Science*. To evaluate our method and test how our classifiers would perform on unseen data, we selected 199 participant entries from prior, uncustomized, implementations of *Land Science*. We took these entries from *uncustomized* versions of the activity the teacher in this study worked to customize. In this case, the customizations to this activity's notebook requirements and assessment criteria, as defined by the core concepts, were not drastically different from the requirements and criteria of the original activity. Thus, this situation provided a case where we could test our classifiers on data that was expected to contain some distribution of the core concepts. In general, however, our method for generating classifiers is meant to accommodate both small customizations, such as we have here, and more drastic ones, such as a case where a teacher creates an entirely new activity. Therefore, we cannot always expect to have such similar data for testing.

The 199 participant entries were manually coded for each core concept by two raters. Both raters had worked with the teacher in this study to define the core concepts and had extensive prior experience coding notebook entries from *Land Science*. Using the process of *social moderation* [6], the raters agreed on the presence or absence of each core concept for each of the 199 entries. From Table 1, we see that the distributions of some concepts are balanced (C2), while others are skewed (C5). However, because we built classifiers based on the textual features of teacher samples, skewness should have a small effect on the performance of the model.

Table 1. Distribution of concepts in data set

Concept	Notations	#Concepts	%Concepts
land use changes	C1	141	72.860
original land use configuration	C2	114	57.280
location of land use change	C3	79	39.690
indicator changes	C4	128	64.320
stakeholder demands	C5	46	23.110

4.2 Threshold Initialization Method

To derive a similarity score threshold, which is needed for the semantic similarity based classifiers, we calculated the similarity scores between the tagged chunks of text for each core concept in the teacher provided examples. Next, we calculated the average and standard deviation of these scores and set our threshold as the average similarity minus one standard deviation for each core concept. The values we obtained using this approach are reported in Table 2, where the last column is the derived threshold for each classifier. Table 2 shows thresholds for both LSA based similarity and the NN based model.

Phrase similarity based on LSA relies on the combination of constituent words a phrases. Hence the similarity score will be more biased towards phrases having common words. While the NN based semantic similarity method [7, 17] projects the phrase pairs into common low dimensional space hence the similarity score obtained will be more consistent irrespective of the presence of common words in the phrases.

Table 2. Derived threshold for LSA based and NN based similarity method

Classifier		Avg.	Std.	Avg. - Std.
C1	LSA	0.584516	0.228474	0.356042
	NN	0.437065	0.122893	0.314172
C2	LSA	0.239488	0.189726	0.049762
	NN	0.242053	0.168682	0.073372
C3	LSA	0.696795	0.103681	0.593114
	NN	0.523347	0.077424	0.445923
C4	LSA	0.278877	0.170271	0.108607
	NN	0.174579	0.124677	0.049902
C5	LSA	0.466482	0.196369	0.270113
	NN	0.149499	0.096005	0.053494

Note: **Avg.**=average similarity score, **Std**=standard deviation.

In Table 2 it is also observed that the standard deviations of similarity scores for NN based models are less than that of the LSA based semantic similarity model in all the five classifiers. This validates our previous understanding that LSA based similarity measures is more biased towards phrases with high degree of word overlap and gives lower score for the phrases with lower degree of or word overlap, resulting high variation in the score. On the other hand, NN based method does not suffer from such biasedness.

4.3 Results

We now present precision and recall results for LSA based and NN based models for the derived thresholds presented earlier and for ideal thresholds (described next). Afterward, we present results for the RegEx based classifiers.

As an alternative to deriving classifiers based on teacher-specified input, we wanted to see how well our methods performed when trained on actual, participant data. That is, when the threshold used in the classifiers to make the final decision was fit based on actual participant data. We call such participant data-trained threshold, the ideal threshold. This ideal threshold could only be computed when participant data is available, which is a major constraint when developing a new internship, as we pointed out earlier.

Figure 1 and 2 shows the precision and recall plot for increasing thresholds of LSA based and NN based similarity methods. These plots were obtained by comparing the model classifications to the manual classifications on the 199 participant entries. It is generally seen that whenever precision increases at a particular threshold, the recall decreases or vice versa. The point of intersection of the precision and recall for a particular classifier gives the ideal precision and recall—that is, the classifier has

balanced performance in terms of precision and recall. From the figure, it is clear that if we want fewer false negatives, for example, the value of the threshold should be increased. In such a case, the precision will be compromised. Therefore, the threshold should be chosen carefully not to compromise either precision or recall to an undesirable extent.

The results obtained with ideal and derived thresholds are summarized in Table 3. These data suggest that, for the ideal thresholds, the LSA based classifiers for core concepts C1 through C4 performed well with the lowest precision and recall value being 0.72. However, the NN based classifiers outperformed the LSA classifiers for all core concepts other than C2. LSA based models depend on the overlapping content words in phrases and the performance suffers in cases where the phrases contain out of vocabulary words. Out of vocabulary here means the LSA similarity relies on pre-built vocabulary from a large corpus that does not contain some of the words, such as proper nouns that are specific to Land Science. However, NN based similarity models rely on letter trigrams from a very large corpus, and every input phrase is converted to letter trigrams. Therefore, the NN based models are capable of capturing the semantics even when there are out of vocabulary words in the phrases or context of the phrases. Hence, the NN based classifiers are superior for these concepts. However, for C2, the NN based classifier lagged in performance by 2% in precision and recall compared to the LSA based classifier because the teacher samples used for C2 contained only short phrases with very few context words and some of the overlapping words in the phrases boosted LSA based classifiers. The classifier C5 performed poorly for both LSA and NN based classifiers.

Table 3. Precision and recall for ideal and derived thresholds for LSA based and NN based similarity method

Classifier		Threshold		Precision		Recall	
		I	**D**	**I**	**D**	**I**	**D**
C1	LSA	0.36	0.35	0.84	0.82	0.84	0.86
	NN	0.34	0.31	0.86	0.84	0.86	0.92
C2	LSA	0.80	0.05	0.80	0.57	0.80	1.00
	NN	0.52	0.07	0.78	0.57	0.78	1.00
C3	LSA	0.38	0.59	0.82	0.92	0.82	0.80
	NN	0.36	0.44	0.86	0.96	0.86	0.78
C4	LSA	0.56	0.11	0.72	0.64	0.72	1.00
	NN	0.46	0.05	0.74	0.64	0.74	1.00
C5	LSA	1.00	0.27	0	0.22	0	0.98
	NN	0.80	0.05	0	0.23	0	1.00

Note: **I**=ideal, **D**=derived.

For the LSA based classifiers, the highest precision using derived thresholds was 0.92 with recall of 0.80 for C3 and the lowest precision was 0.22 with recall of 0.98 for C5. As we saw with the derived thresholds, NN based classifiers generally outperformed their LSA based classifiers counterparts, with the exception of the recall for concept C3

The results in Table 3 suggest that a good threshold could be derived without participants' data. The high recall and precision using derived thresholds for concepts C1 and C3 suggest the possibility of assessing the core concepts in participant notebook entries with classifiers generated using only the teacher's sample responses. However, when compared to the results using the ideal thresholds, classifiers C2, C4 and C5 did not perform well; their derived thresholds differed largely from their ideal thresholds, and their precision and recall suffered. The relatively low derived threshold values for these concepts suggests that their associated examples, which were used to calculated the thresholds, were semantically dissimilar. Dissimilar examples for a given concept could imply an ill-defined concept and that the provided examples do not represent it well. Alternatively, dissimilar examples could imply a complex or varied concept that requires highly different examples to represent it fully. Because we cannot distinguish between these cases automatically, we plan in future work to set a best guess threshold of 0.5 in such cases.

Table 4. Performance of regular expression model

Concepts	Precision	Recall
C1	0.963	0.551
C2	0.640	1.000
C3	1.000	0.746
C4	0.791	0.890
C5	0.894	0.739

Table 4 shows the precision and recall of RegEx based classifiers. Here the performance for concepts C2, C4, and C5 is more interesting when we compare those values with the previously discussed result. For example, the precision and recall for C5 improved impressively with values 0.89 and 0.73 respectively, whereas in previous case those values were either undefined or 0 precision with recall 1. Furthermore, the precisions of C1 and C3 are high, however the recalls are relatively low. Qualitatively investigating these results suggested that participants entries expressed these concepts in a variety of ways that were not captured by the regular expression lists.

Given that we see improvements for some core concepts using the regular expression based approach, these results suggest that the teacher provided samples on which the similarity measures where based may not have included a variety of key terms that could indicate the presence or absence of these core concepts. Comparing the sample responses and the keywords provided revealed that the samples indeed did not contain many of the keywords in the list. In some cases, the keywords were synonyms or other instances of particular kinds of words provided in the sample responses. For example, in *Land Science*, there are sixteen stakeholders who give demands on zoning plans. The core concept C5, *stakeholder demands*, is meant to capture references to these 16 stakeholders in participant notebook entries. Examining the teacher provided samples, we found that only four stakeholders were covered, while the keyword list for the core concept mentioned all sixteen. We plan in future experiments to either ask teachers to provide enough samples to cover finite sets of semantic content such as this or to incorporate the provided keyword list into the semantic similarity methods as extra samples.

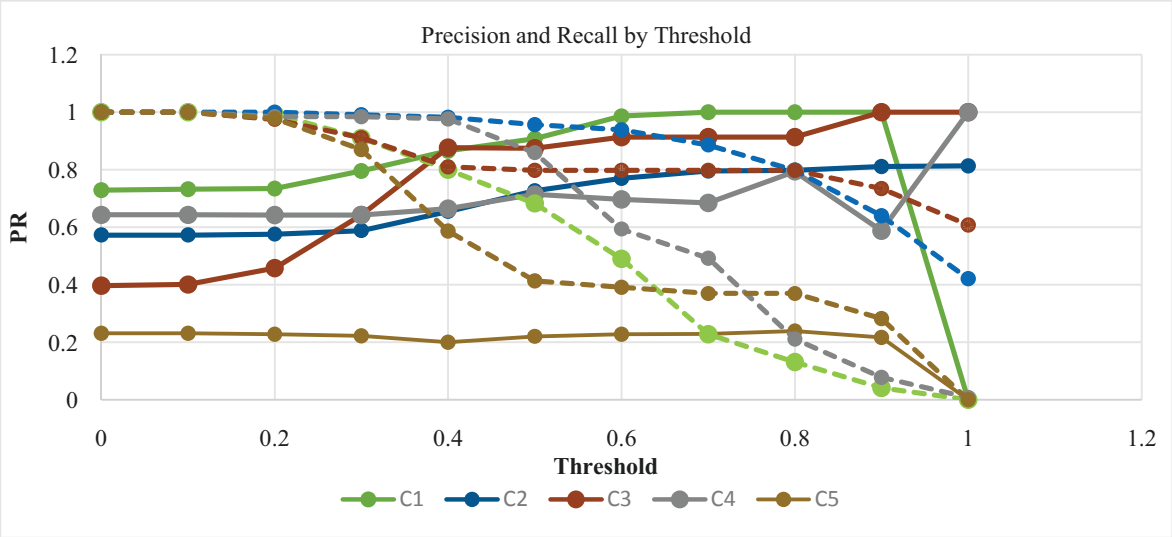

Figure 1. Precision and recall for LSA based similarity thresholds (solid lines are precision; dotted lines are recall)

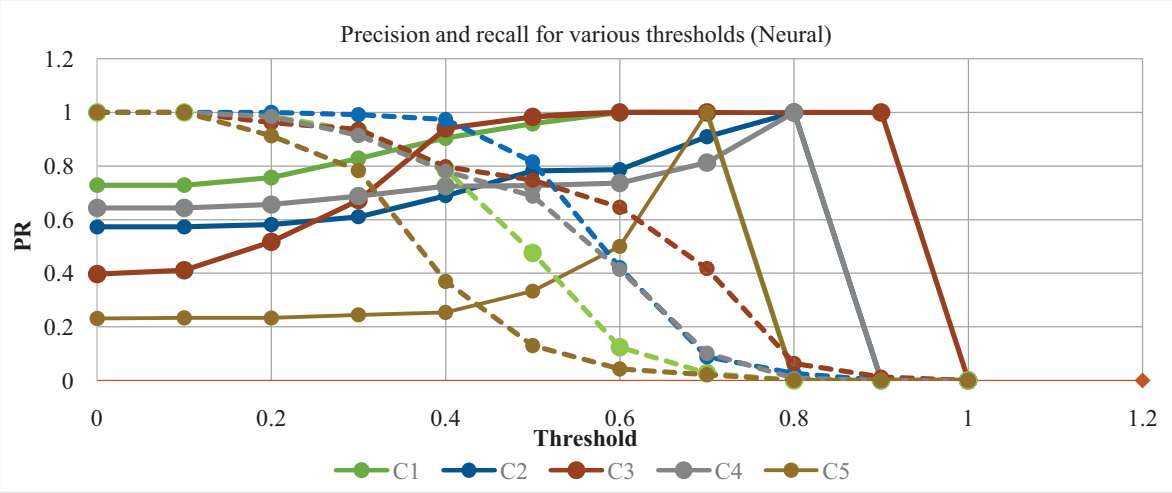

Figure 2. Precision and recall for neural network based similarity thresholds (solid lines are precision; dotted lines are recall)

5. CONCLUSIONS

In this paper, we investigated a method for creating classifiers for virtual internship notebook entries using teacher provided specifications without the use of participant data. Our classifiers used LSA based and NN based semantic similarity methods to capture the general semantic relationships among concepts. We also investigated regular expression based classifiers. The results are impressive in the sense that some classifiers, using both LSA and NN, gave high precision and recall values using thresholds derived without participant data, which suggests that our general method is plausible.

Furthermore, the superiority of the NN classifiers over the LSA classifiers suggests that NN methods are preferable when the participant responses vary widely in terms of style, content, and word overlaps with the teacher provided sample response.

The improved performance for some core concepts, such as C5, using regular expression based classifiers implies that such classifiers performed better for concepts whose sample responses did not contain a variety of keywords, despite the benefits we saw for NN models. These results suggest that, in some cases, teachers may need to provide more exhaustive samples, and that provided keywords and regular expression based classifiers may supplement a semantic similarity approach.

In future work, we will investigate a method to combine the classifiers in order to better understand how performance of one model is boosted by another in the scenario where participants responses vary widely compared to the sample responses. We will also see how the performance be affected by setting up the thresholds to 0.5 for concepts C2, C4 and C5.

Our work has several limitations; most obviously, we used participant data in to evaluate the performance of some of our classifiers. In the real use case of our method, we cannot expect to have such data available. We want to make clear, however, that our purpose in using participant data was not to train better classifiers, but to evaluate our method for generating them. Thus, our results suggest that this method can produce classifiers that would perform well on unseen data, but more refinements are needed.

6. ACKNOWLEDGMENTS

This work was funded in part by the National Science Foundation (DRL-0918409, DRL-0946372, DRL-1247262, DRL-1418288, DUE-0919347, DUE-1225885, EEC-1232656, EEC-1340402, REC-0347000), the MacArthur Foundation, the Spencer Foundation, the Wisconsin Alumni Research Foundation, and the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison. The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.

7. REFERENCES

[1] Cai, Z., Graesser, A. C., Forsyth, C., Burkett, C., Millis, K., Wallace, P., ... & Butler, H. (2011). Trialog in ARIES: User input assessment in an intelligent tutoring system. In *Proceedings of the 3rd IEEE international conference on intelligent computing and intelligent systems* (pp. 429-433).

[2] Corley, C., & Mihalcea, R. (2005). Measuring the Semantic Similarity of Texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. Ann Arbor, MI.

[3] Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment, 5*(1).

[4] Fernando, S. & Stevenson, M. (2008). A semantic similarity approach to paraphrase detection, In *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics* (pp. 45-52).

[5] Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36, 2(2004), 193-202

[6] Herrenkohl, L. R., & Cornelius, L. (2013). Investigating elementary students' scientific and historical argumentation. *Journal of the Learning Sciences*, *22*(3), 413–461.

[7] Huang, P. S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. (2013, October). Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 2333-2338). ACM.

[8] Leacock, C., and Chodorow, M. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389-405.

[9] Lintean, M. C., & Rus, V. (2012, May). Measuring Semantic Similarity in Short Texts through Greedy Pairing and Word Semantics. *In FLAIRS Conference*.

[10] Rus, V., Gautam, D., Swieki, Z., & Shaffer, D. W. (2016, June). Assessing Student-Generated Design Justifications in Virtual Engineering Internships. In *Educational Data Mining* 2016.

[11] Rus, V., Lintean M., Graesser, A.C., & McNamara, D.S. (2009). Assessing Student Paraphrases Using Lexical Semantics and Word Weighting. *In Proceedings of the 14th International Conference on Artificial Intelligence in Education*, Brighton, UK.

[12] Rus, V., Lintean, M. C., Banjade, R., Niraula, N. B., & Stefanescu, D. (2013, August). SEMILAR: The Semantic Similarity Toolkit. In *ACL (Conference System Demonstrations)* (pp. 163-168).

[13] Salton, G., and Lesk, M. 1971. Computer evaluation of indexing and text processing. Prentice Hall, Ing. Englewood Cliffs, New Jersey. 143–180.

[14] Shaffer, D. W. (2006). *How Computer Games Help Children Learn*. Macmillan.

[15] Shaffer, D. W., Ruis, A. R., & Graesser, A. C. (2015). Authoring Networked Learner Models in ComplexDomains. In *Design recommendations for intelligent tutoring systems*, 179.

[16] Shermis, M.D. & Burstein, J. (2003). *Automated Essay Scoring: A Cross Disciplinary Perspective*. Lawrence Erlbaum Associates, Mahwah (2003).

[17] Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014, November). A latent semantic model with convolutional-pooling structure for information retrieval. *In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 101-110). ACM.

[18] Swiecki, Z., Midsfelt, M., Stoddard, J., Shaffer, D.W., (in press). Dependency-Centered Design as an Approach to Pedagogical Authoring. InBaek, Y. (Ed.) *Game-Based Learning: Theory Strategies and Performance Outcomes*. Hauppauge, NY: NOVA.