Annotation-free Automatic Examination Essay Feedback Generation

Filipe Altoe
Department of Computer Science
Georgia Institute of Technology
Atlanta, United States
faltoe3@gatech.edu

David Joyner
Department of Computer Science
Georgia Institute of Technology
Atlanta, United States
djoyner3@gatech.edu

***Abstract*— Many of the demands of scale appear at odds with quality, such as in the case of online MOOC degree programs. This paper presents research that focuses on reducing this dichotomy. Examination essays are important learning tools in for-credit courses. Psychology studies show they drive deeper learning strategies and invoke higher levels of cognitive processing in students. High-quality feedback is another critical aspect of students' learning in advanced education. Key components of high-quality feedback are how expeditiously feedback is presented and how unambiguous it is. Examination essays are time-consuming for educators to grade, and at scale this may create an incentive for superficial feedback that may also reach students in a less than ideal turnaround time. This imposes clear scalability constraints and a potential decrease in course quality in online for-credit MOOC programs. We propose an annotation-free artificial intelligence-based approach for the automatic generation of examination essay rubrics and its subsequent utilization as part of high-quality feedback to students. It utilizes a combination of the natural language processing techniques TextRank and semantic similarity for automatic rubric generation and concept map creation for feedback presentation. The feedback is immediately available upon essay submission and offers the student a conceptual analysis of the submitted essay against the assignment's learning objectives.**

Keywords—scalability of MOOC degree programs, high-quality feedback on MOOCs, automatic rubric generation, TextRank Semantic Similarity, natural language processing

I. Introduction

Essay assignment examination is a fundamental tool in education. Research has shown they drive deeper learning strategies and invoke a higher level of cognitive processing in students [1].

Two different types of essays are common in modern education; examination essays and essay tests. Examination essays differ from essay tests insomuch that it extends beyond answering a specific question. It includes answering an open question through content knowledge and its understanding rather than solely composition skills and factual knowledge. An example of an examination essay question could be: "*Defend whether it is possible for European companies that focus on personal data to continue in business with the advent of the GDPR*". Note that this question requires not only a deep understanding of GDPR itself, but also probes for an impact analysis on how GDPR can potentially affect European companies. An example of an essay test question on the same general theme could be: "*Describe what GDPR says about usage of personal data.*". This essay test question probes the student for factual knowledge about the content, while not necessarily requiring a deep level of understanding on how this knowledge can be applied in real-world situations.

Lately, educators have been favoring examination essays as they don't focus on memory for facts and formulas [2], but on deeper demonstration of knowledge by the students. Arguably, essay tests are easier for scorers to grade as their focus is on rhetoric, grammar and style, and not as much on expert content knowledge. Examination essays require a deeper level of cognitive attention and potential need for abstract thinking by the grader. Furthermore, graders of examination essays are required to have a much deeper level of domain knowledge than graders of test essays. These differences may lead to superficial and long turnaround time for feedback to be provided to students.

At scale, the extended number of students enrolled in for-credit MOOC courses in comparison to their on-campus counterparts aggravate the issue. Some for-credit MOOC courses may command a viable maximum number of teaching assistants to be engaged. This increased ratio of students per grader may cause the quality of feedback to worsen, directly impacting the overall quality of the MOOC program.

Psychology research has shown that feedback to students improves their learning [3]. However, for feedback to be effective, it requires the following attributes [4]:

1) Clear and unambiguous. Students may not understand how to improve their work based on such comments [5];
2) Constructive. Negative feedback may act as a demotivator for students who may already have low self-esteem [6];
3) Expeditious. [7] correlates the fast and high-quality feedback to its relevancy.

High-quality feedback for MOOC courses that leverage examination essays is an important topic of research and is the main focus of this work. The reemergence of Artificial Intelligence (AI) in the last few years drove focused research to applications of AI-based essay auto-graders, the so-called Automated Essay Scoring (AES) systems, as a potential solution to the scalability and turnaround time issues stated above.

As techniques of Natural Language Processing (NLP) evolved in the last few years, the research expanded the field of AES into one considering the understanding of the concepts presented in the essay [8]. This expansion also brought the possibility for the application of automated tools to examinations essay[9], not only to essay tests.

978-1-7281-2549-7/19/$31.00 ©2019 IEEE 23-25 October 2019, Milwaukee, WI, USA

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on July 14,2020 at 18:59:44 UTC from IEEE Xplore. Restrictions apply.

However, the seminal AES approaches listed in [10], as well as more modern approaches such as [11], [12] and [13] focus on the grading task itself. They either do not provide feedback to students altogether or treat it as a byproduct of the grading activity.

The approaches that are more promising in support of feedback to students such as [14], [15], [16], [17], [18] and [19] have two relevant drawbacks: 1) are machine learning-driven and consequently require a significant number of training essays; 2) the essays need to be annotated by a human grader; which usually demands substantial time investment by qualified graders. These two combined characteristics may prevent these approaches from removing the scalability constraints that automated tools could help address.

Few AES approaches include feedback to students as their main artifact. Furthermore, the ones that focus on it invariably require a large number of training essays. Lastly, to the best of our knowledge, the literature does not include annotation free approaches for automatic generation of examination essay rubric and its consequent utilization for generating feedback to students.

We propose an approach for the automatic creation of examination essay rubrics that: 1) does not require a large training set of essays; 2) does not require human essay annotation and 3) could be ultimately used as a basis for automatic generation of high-quality feedback to students.

As a first contribution to the field, this work presents an NLP-based annotation-free approach for the automatic generation of rubrics that combines TextRank [20] and Automatic Concept Map Generation [21], that, to the best of our knowledge is novel. Furthermore, feedback on examination essays is provided to students through the analysis of conceptual differences between the automatically generated rubric and the student submitted essay.

II. Related Work

Project Essay Grader (PEG) was developed by Ellis Page in 1966 [22], and it is considered to be the seminal work on AES. It uses statistical correlation to try and replicate essay human generated grades. Its architecture has been improved over the years, and its more modern incarnation includes various dictionaries and classification schemes [10]. However, it completely ignores content-related features, focusing mainly on style, therefore, it doesn't provide any feedback to students or analyze content.

IEA utilizes a semantic text-analysis method named Latent Semantic Analysis [23], or LSA. LSA is defined as a statistical model that allows comparison of semantic similarity of text [24]. Once trained, the essay is represented as an LSA vector [25] and then compared to other texts from the corpus. IEA leverages LSA to focus more on the context related features rather than the form related ones, however, the approach still doesn't offer feedback to students on semantics and overall concepts.

E-rater is an AES corpus-based approach that leverages Natural Language Processing (NLP). Usually, when this type of approach is utilized on AES, the availability of annotated corpus text is required. However, E-rater doesn't require an annotated corpus, which is a clear advantage over other tools that use the same approach [11].This AES approach, however, doesn't provide any feedback to students.

The Bayesian Essay Test Scoring System (BETSY), as the name suggests, employs Bayesian theorem to find the most likely classification of an essay into a four-point nominal scale (extensive, essential, partial, unsatisfactory) using a large set of features including both content and style [14]. It is claimed that BETSY includes the best features of PEG, LSA and E-Rater, along with its own inherent benefits. However, it doesn't focus on feedback generation either.

[26] states that an AES system, to be accepted and used without human interference, needs to include deep semantic analysis. It proposes a novel approach that detects semantic errors and provides semantic essay feedback to the student. However, it doesn't provide content feedback, a necessity according to the definition of high-quality feedback presented herein.

Neural networks (NN) and Deep learning (DL) are arguably two of the most explored topics in modern NLP. [15] proposes an approach based on recurrent NN applied to the task of AES. It presents the analysis of several different NN models, showing overall good performance in relation to other open-source documented approaches, but it focuses on auto-grading alone.

[16] focuses on an ML approach in combination with a domain-based ontology. The use of ontology allows for another dimension of evaluation stretching beyond keyword presence and context of keyword, but also whether the keywords are appearing in the right context. However, its weakness lies in the fact that it doesn't provide students with feedback about missing concepts in their essays.

In [27], the focus of the research is in using machine learning techniques to analyze the structure of the essay explanations, the connections between causes and effects. It focuses the analysis on explanatory essays; carrying the promise of improving the quality of feedback to students and attempting to create a causal concept relationship.

[19] proposes a Knowledgebase AI-based approach to AES that is similar to hierarchical classification. It reports a 55% exact accuracy between predicted essay scores and human scores. Though such results can be seen as promising, the authors go on to conclude the work by stating that there are some disadvantages to using a hierarchical approach and that its main strength lies on the potential to provide good feedback to the writer, not necessarily on the essay grading. However, the approach doesn't provide evidence for this claim.

[18] offers a cognitive-based system applied to AEE tasks. It is one that approaches a full cognitive architecture the most in the literature; including a sensory acquisition module, a score analyzer module and a background knowledge constructor model. The architecture is claimed to be able to provide individualized feedback to students. However, the work lacks in analysis of a higher number of test cases; rendering its results inconclusive.

Focusing the research on examination essays, [17] proposes a machine learning approach that attempts to connect cause to effect of the student writing. However, it does provide two important drawbacks. 1) As most machine learning-based approaches, it requires a significant number of training essays; 2) It requires the training essays to be

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on July 14,2020 at 18:59:44 UTC from IEEE Xplore. Restrictions apply.

annotated by a human grader; which may require a significant time investment, depending on the number of training essays needed by the model and the complexity of the essay theme.

Though it seems clear that Artificial Intelligence is making good strides to improve the overall effectiveness of auto-graders, to the best of our knowledge, few approaches include feedback to students as its main artifact. Furthermore, the approaches that do add student feedback, invariably don't cover examination essays. Lastly, to the best of our knowledge, the literature doesn't include annotation free approaches for the automatic generation of examination essay rubric and its consequent utilization focused on the generation of feedback to students.

III. Methodology

The central idea of the proposed approach is the comparison of the content of a student's response to a given open-ended question to an ideal answer that is considered to include all required concepts as dictated by the assignment's learning objectives. This ideal answer can be considered to be a rubric to the given assignment.

One possible path to this comparison would be the generation of such rubric essay by the teaching body. It would be an essay written in a manner to cover all required concepts and their relationships, in alignment to the learning goals.

There are some drawbacks to this approach. A notable one is, if a single person is responsible for the creation of the golden response, there is a missed opportunity for various perspectives to be utilized in the generation of such golden response.

An approach that can mitigate that drawback is the utilization of essays from past semesters, created by students and identified by the teaching body as exemplary. Exemplary, by definition, is the representation of the best of its kind. Thus, exemplary essays can be seen as the set of essays, produced by several students, that were selected as containing all or the majority of concepts and their relationships required to capture the learning objectives of the assignment.

This approach presents the extra advantage of involving several graders in the process of identification of the exemplary assignments. Therefore, potential individual grader bias and halo effects [28] that may have swayed an essay to have been included as part of the exemplary pool can be filtered out through the auto-generation of a rubric that best capture the intersection of the most relevant concepts of the pool.

For these reasons, the proposed approach of automated generation of rubric leverages the utilization of exemplary examination essays as its training set. The chief concepts are the execution of concept summarization of all exemplary assignments and consequent similarity analysis between them. Arguably, the exemplary summary that carries the highest score of similarity to all other summaries can be assumed to be the most conceptually complete assignment - it is the one that includes concepts that are also present in all the other summaries. Therefore, this exemplary summary is elected the rubric summary.

In general terms, a tool to automatically generates rubrics using as a basis a set of exemplary examination essays shall contain the following top-level components: identification of essay concepts, ranking of the main concepts, summarization of essays, similarity analysis, rubric generation and rubric visualization.

TextRank [20] is the NLP approach that best fits the identification of concepts and summarization tasks. TextRank not only identifies the main concepts from free text, but it also ranks them in terms of importance. Ranking provides an extra facility for the study of a potential optimal size of the rubric summary for a given assignment. This size is determined via the maximum number of words and phrases.

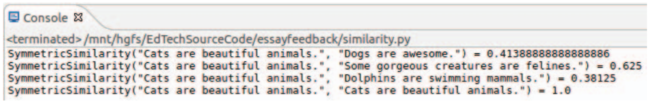
The PyTextRank [29] Python library is being leveraged for the implementation of TextRank. It calculates a significance weight for each sentence, using MinHash to approximate a Jaccard distance from key phrases, as determined by the TextRank algorithm. PyTextRank allows the creation of JSON files to store the graph representation of the analyzed essay.

Semantic similarity of long texts can be considered a current open problem in the literature. However, since the approach is based on summaries of essays, the problem is reduced to similarity between short texts.

Python NLTK for PoS tagging and WordNet corpus are leveraged. The algorithm that yielded the best results for calculating similarity between summaries is described as follows:

1) It focuses on one summary and calculates similarity score between the focus summary and every other summary generated by the summarization module;
2) It averages the scores and assigns the value as the similarity score for the focus summary;
3) Repeats 1 and 2 for every summary as a focus summary;
4) The summary with the highest average similarity score is made the rubric summary

Figure 1 shows the scores returned by the similarity algorithm when comparing two simple sentences. These scores are used as benchmarks for the essay summaries similarity analysis.



```
<terminated> /mnt/hgfs/EdTechSourceCode/essayfeedback/similarity.py
SymmetricSimilarity("Cats are beautiful animals.", "Dogs are awesome.") = 0.41388888888888886
SymmetricSimilarity("Cats are beautiful animals.", "Some gorgeous creatures are felines.") = 0.625
SymmetricSimilarity("Cats are beautiful animals.", "Dolphins are swimming mammals.") = 0.38125
SymmetricSimilarity("Cats are beautiful animals.", "Cats are beautiful animals.") = 1.0
```

Figure 1. Similarity Score Benchmarks.

.

Visualization is considered a very important sub-task in the scope of the research as it directly affects the quality of the generated feedback to students. Concept maps were selected as the mechanism for visualization of the generated rubric. Automatic Concept Map Generation (ACMG) is another area of active current research by the community, and it is also considered to be an open problem.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on July 14,2020 at 18:59:44 UTC from IEEE Xplore. Restrictions apply.

Even widely utilized ACMG from text tools, such as the Stanford Open Information Extraction Tool [30], don't yield satisfactory results for multi-concept sentences composing short texts.

Therefore, a custom algorithm for automatic concept map generation was developed. This task was facilitated by the fact that the rubric summary is represented in JSON graph format. The algorithm works as follows:

1) Retrieve all concepts that are n-grams from the JSON file to be processed first;
2) From the summary text, retrieve all sentences that include these n-grams;
3) Create a concept map for each one of these sentences
    a. Concepts are created based on each concept PoS tag that is also part of JSON file;
    b. Process all "Noun-Verb" and "Verb-Noun" pairs for each sentence, so multi-concept sentences are processed;
4) Repeat 2 and 3 for the 1-gram concepts;
5) Perform interlinking of concepts that are present in more than one sentence.

IV. Results

Experiments were conducted to validate the proposed approach. The training set used was composed of ten essays identified as exemplary by the teaching body of an Artificial Intelligence course in a for-credit online MS in CS program. A summary of the assignment question asked to the students was:

"Research GDRP passed by the European Union. Select an example of a company/industry for which data personalization is deeply embedded in its model and defend whether it is possible for users in the EU to use the services without waving their GDRP rights."

As it can be seen on the open question above, the student's responses qualify as examination essays. Each student's exemplary response varied from 400 to 650 words in length. The summary size was set to have a maximum of 200 words and ten phrases.

The plot below shows the average similarity score for each of the ten exemplary summary essays generated by the approach.

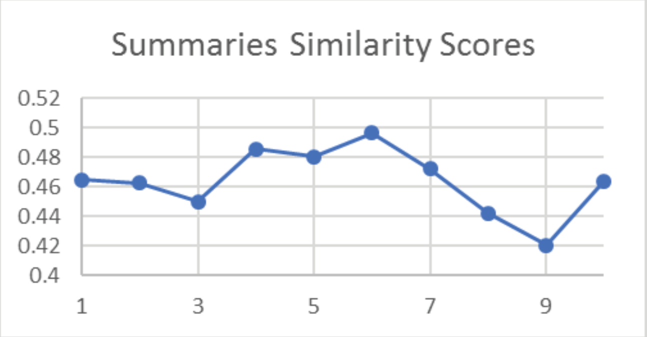


Figure 2. Summaries Similarity Scores.

As it can be seen on the plot, all exemplary summaries present close similarity averages. This indicates a solid criterion was utilized for the selection of an essay as exemplary by this class' teaching body. Summary number 6 was selected as the summary rubric as it shows the highest similarity score of the set. The similarity score of 0.5 is considered an acceptable value based on the benchmark similarity scores used on simple sentences presented in Figure 1. The automatically generated rubric is the automatically generated exemplary summary showing the highest level of concept similarity with all the other auto-generated exemplary summaries. For the example presented herein, the following passage was the auto-generated rubric:

"*Research the recently-passed General Data Protection Regulation passed by the European Union. Describe what the regulation says about the usage of personal data to personalize individual user experiences online. Analyze how that regulation might apply to the use of artificial intelligence to create personalized experiences.*
Then, select an example of a device, company, or industry for which personalization is deeply embedded in its functional purpose or business model. Personalization should be deeply rooted in the purpose or model: Amazon, for example, uses personalized recommendations, but it is not difficult to imagine a user being able to specifically opt out of those recommendations and still use most of the service. Select a device, company, or industry for which, without personalization, there is no service.
Then, evaluate how these devices, sites, or services may be adapted to these GDPR restrictions. Determine and defend whether it is even possible to allow users in the European Economic Area to use these tools without waiving their GDPR rights."

For feedback generation, the execution of similarity analysis was conducted between the generated rubric and test essays that were not part of the exemplary essays. Figures 3 and 4 illustrate an example of feedback based on a student submitted essay. Figure 3 shows the concept map including all concepts successfully captured in the student's essay that are also part of the auto-generated rubric. Figure 4 completes the feedback to the student with the presentation of the concepts that were included in the rubric but missed in the student's essay.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on July 14,2020 at 18:59:44 UTC from IEEE Xplore. Restrictions apply.

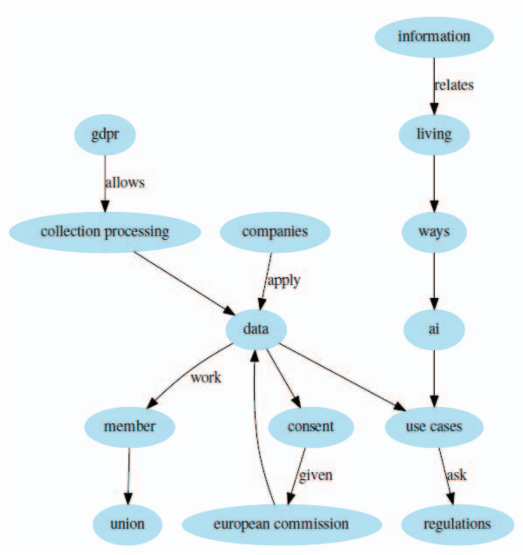


Figure 3. Feedback - Included Concepts

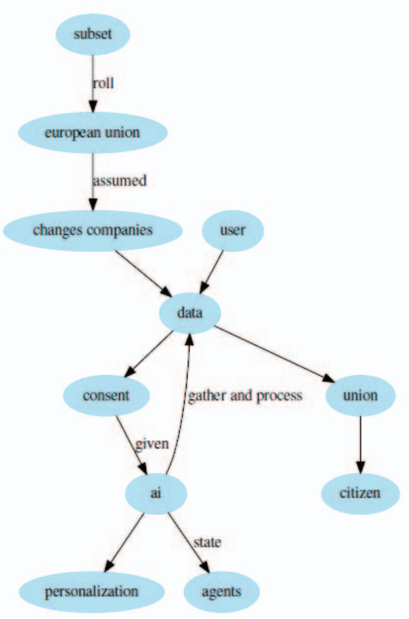


Figure 4. Feedback – Missing Concepts (right)

In summary, the workflow for automatic feedback generation of examination essays proposed is as follows:

1) Examination essay question is posed to students;
2) Students submit essays;
3) Teaching body selects essays to be part of exemplary pool;
4) Our approach takes the pool of exemplary essays as input and creates a rubric through: a) Summarization via ranking of the most relevant concepts on each essay; b) Concept similarity analysis of each summary with all the other generated summaries; c) Selection of the summary with highest similarity value as the rubric;
5) Generation of feedback to students presented in the form of two concept maps: one showing all concepts presented in the student's essay that are also included in the rubric; and the second showing all concepts present in the student's essay that are not included in the rubric.

V. Conclusions and Future Work

Essay assignment examination is a fundamental tool in education, especially in higher education. Research has shown they drive deeper learning strategies and invoke a higher level of cognitive processing in students.

Psychology research has shown that feedback to students improves their learning. High-quality, effective feedback is required to be unambiguous, constructive and expeditious. Examination essays require a deeper level of cognitive attention and potential need of abstract thinking by the grader. Furthermore, graders for examination essays are required to have a much deeper level of domain knowledge than graders of test essays. These differences may lead to superficial and long turnaround time for feedback to be provided to students.

At scale, the extended number of students enrolled in for-credit MOOC courses in comparison to their on-campus counterparts aggravate the issue. Some for-credit MOOC courses may command a viable maximum number of teaching assistants to be engaged. This increased ratio of students per grader may lead the quality of feedback to worsen, directly impacting the overall quality of the MOOC program.

This work proposes an artificial intelligence-based approach for the generation of high-quality examination essay feedback to students. It aimed to mitigate two typical problems present in machine learning-based approaches to auto-graders: 1) require a significant number of training essays; 2) essays need to be annotated by a human grader; which usually demands substantial time investment by qualified graders.

It presents a novel annotation free approach for the automatic generation of examination essay rubrics from a small set of pre-graded exemplary essays. It leverages the natural language processing techniques of TextRank and Automatic Concept Map Generation for automatic rubric generation and creation of immediate high-quality feedback to students. Feedback is based on conceptual similarities and differences between the automatically generated rubric and student' submitted essays.

The utilization of exemplary essays as a training set for the proposed approach allows automatic rubric generation based on a small training set and removes the need for human essay annotation. The same concept summarization/ranking process can be applied to the student's submitted essay for subsequent feedback generation. Concept map generation and classification of all student's essay concepts into missing and included concepts provide visual and immediate high-quality feedback to students on their submitted examination essays.

Future work includes the following opportunities:

1) Analysis of the proposed approach on examination essays with a higher number of words and different themes;
2) Experiments with a large set of students' examination essays across multiple semesters;

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on July 14,2020 at 18:59:44 UTC from IEEE Xplore. Restrictions apply.

3) Analysis of similarity score on a larger number of exemplary essays;
4) Research to determine whether the rubric summary size can be made a function of the average size of exemplary essays, instead of a constant value.

REFERENCES

[1] K. Scouller, "The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay," *Higher Education,* vol. 35, no. 4, pp. 453-472, 1998

[2] M. Freeman and J. McKenzie, "large classes: the case for an online self and," *Peer learning in higher education: Learning from & with each other,* p. 156, 2001K. Scouller, "The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay," *Higher Education,* vol. 35, no. 4, pp. 453-472, 1998.

[3] N. R. Council, R. R. Cocking, A. L. Brown, and J. D. Bransford, "How people learn: Brain, mind, experience, and school," 1999: National academy.

[4] M. Taras, "Using assessment for learning and learning from assessment," *Assessment & Evaluation in Higher Education,* vol. 27, no. 6, pp. 501-510, 2002

[5] D. Hounsell, "Student feedback, learning and development," *Higher education and the lifecourse,* pp. 67-78, 2003.

[6] P. Young, "'I might as well give up': Self-esteem and mature students' feelings about feedback on assignments," *Journal of Further and Higher education,* vol. 24, no. 3, pp. 409-418, 2000.

[7] S. Brown and P. Knight, *Assessing learners in higher education*. Routledge, 2012.

[8] M. D. Shermis, J. Burstein, and S. A. Bursky, "Introduction to automated essay evaluation," in *Handbook of automated essay evaluation*: Routledge, 2013, pp. 23-37.

[9] D. Horowitz, "Essay examination prompts and the teaching of academic writing," *English for Specific Purposes,* vol. 5, no. 2, pp. 107-120, 1986.

[10] S. Valenti, F. Neri, and A. Cucchiarelli, "An overview of current research on automated essay grading," *Journal of Information Technology Education: Research,* vol. 2, pp. 319-330, 2003.

[11] Y. Attali and J. Burstein, "Automated essay scoring with e-rater® V. 2," *The Journal of Technology, Learning and Assessment,* vol. 4, no. 3, 2006.

[12] P. W. Foltz, D. Laham, and T. K. Landauer, "The intelligent essay assessor: Applications to educational technology," *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning,* vol. 1, no. 2, pp. 939-944, 1999.

[13] S. Elliot, "IntelliMetric: From here to validity," *Automated essay scoring: A cross-disciplinary perspective,* pp. 71-86, 2003.

[14] L. M. Rudner and T. Liang, "Automated essay scoring using Bayes' theorem," *The Journal of Technology, Learning and Assessment,* vol. 1, no. 2, 2002.

[15] K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1882-1891.

[16] M. S. Devi and H. Mittal, "Machine learning techniques with ontology for subjective answer evaluation," *arXiv preprint arXiv:1605.02442,* 2016.

[17] P. Hastings, S. Hughes, and M. A. Britt, "Active Learning for Improving Machine Learning of Student Explanatory Essays," in *International Conference on Artificial Intelligence in Education*, 2018, pp. 140-153: Springer.

[18] L. Li and V. Sugumaran, "A cognitive-based AES model towards learning written English," *Journal of Ambient Intelligence and Humanized Computing,* pp. 1-10, 2018.

[19] D. S. McNamara, S. A. Crossley, R. D. Roscoe, L. K. Allen, and J. Dai, "A hierarchical classification approach to automated essay scoring," *Assessing Writing,* vol. 23, pp. 35-59, 2015.

[20] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

[21] C. Aguiar, "Concept maps mining for text summarization," Universidade Federal do Espírito Santo, 2017.

[22] E. B. Page, "The imminence of... grading essays by computer," *The Phi Delta Kappan,* vol. 47, no. 5, pp. 238-243, 1966

[23] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science,* vol. 41, no. 6, pp. 391-407, 1990.

[24] P. W. Foltz, "Latent semantic analysis for text-based research," *Behavior Research Methods, Instruments, & Computers,* vol. 28, no. 2, pp. 197-202, 1996.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research,* vol. 3, no. Jan, pp. 993-1022, 2003.

[26] K. Zupanc and Z. Bosnić, "Automated essay evaluation with semantic analysis," *Knowledge-Based Systems,* vol. 120, pp. 118-132, 2017.

[27] P. Hastings, S. Hughes, and M. A. Britt, "Active Learning for Improving Machine Learning of Student Explanatory Essays," in *Artificial Intelligence in Education*(Lecture Notes in Computer Science, 2018, pp. 140-153.

[28] R. E. Nisbett and T. D. Wilson, "The halo effect: evidence for unconscious alteration of judgments," *Journal of personality and social psychology,* vol. 35, no. 4, p. 250, 1977.

[29] P. Nathan, "PyTextRank, a Python implementation of TextRank for text document NLP parsing and summarization," ed. https://github.com/ceteri/pytextrank/, 2016.

[30] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55-6.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on July 14,2020 at 18:59:44 UTC from IEEE Xplore. Restrictions apply.