An overview of an automated essay grading systems on content and non content based

Ramesh Dadi¹, Syed Nawaz Pasha ², Mohammad Sallauddin³, Chintuju Sidhardha⁴, A.Harshavardhan⁵

¹ Member center for Artificial intelligence and Deep learning, Computer Science & Engineering, S R Engineering College, Warangal,Telangana,India
², ³, ⁵Computer Science & Engineering, S R University, Warangal,Telangana,India
⁴Sumathi Reddy Institute of Technology for Women, Warangal,Telanagana,India

E-mail: dadiramesh44@gmail.com

Abstract. Assessment is considered to play an essential function inside the Educational System. The interest of using automatic tools by human beings has been increased. So in the same way student response evaluation in education system with automatic assessment systems has grown exponentially in the last couple of years. Due to the increasing number of students and the use of online MOOC courses and lack of time and lack of consistency, assessment is shifted to automatic assessment. In this regard many of the researchers worked on it to make the assessment process easy. And they succeeded in assessment objective-type questions: i.e. multiple choices. Now here comes an interesting part is to assess the essays with automated tools. In this area, more number of researches worked and invented some tools for grading the essays but not up to the mark. Most of the assessment tools were assigning grades based on the style that is the number of sentences, the number of words, parts of speech, and grammar, etc. And remaining is considering features like semantic analysis, using natural language tools grading the essays.

Keywords: computer based assessment, natural language processing, Machine learning, deep learning.

1. Introduction
   The automated essay grading is a computer based evaluation of student responses. Now a day’s education system is shifted to the online based education system and assessment also automatically. This paper we present a survey on automated essay grading systems and their classification with accuracy. Till now more than 15 essay grading systems were there and their working on two categories like content-based and style based assessment few are on both. But most of the tools working on a style based and statistical based assessment by considering features like number of sentences, number of words, parts of speech and grammar, etc. And remaining is considering features like semantic analysis, using natural language tools grading the essays.

2. overview
   2.1 Project essay grade:
   The first automated scoring system was developed by Page, and it grades the essay based on the structure, not on the content of the essay. The PEG system works on statistical features from essay. The features include like fluency, finding of essay length, preposition counting, POS, grammar, relative pronouns, etc. And the intrinsic variable is another parameter. And this variable works on interest within the essay to simulate human rater grading. But there is no distinction between proxes and intrinsic variable both provides a theoretical justification for the quantitative approach.

   And PEG system calculates the grade of an essay based on a statistical approach like a
linear regression model. And the model will be trained based manually previously marked essays from this it finds regression coefficient to assign a score for unmarked essays. Its works on only statistical based features no content based features retrieved.

2.2 Intelligent essay assessor:
IEA was proposed by Hearst, Jerrams-Smith, Soh, and Callear. It uses techniques called Latent Semantic Analysis (LSA) [9], Topic modeling, Singular Value Decomposition. It considers both structural and the context in grading an essay Topic Modeling is the first concept used to find the hidden themes from the given document and it discovers a group of words that are co-occurring in text documents using the UN-supervised learning technique [10]. LSA [9] uses the bag of words into the term documentation matrix that represents occurrences of terms in a document. The LSA uses singular value decomposition on the documentation matrix to subdivide the document matrix into three matrices with different dimensions that reproduce the original matrix after multiplying all 3 separated matrices. The reduced dimension of these three matrices in which the word-context associations can be represented, new relationships between words and context are induced when reconstructing new matrix from close approximations to the original matrix from the reduced Dimension matrices. These new relationships are made manifest, whereas before the SVD they were hidden or latent. LSA is not guaranteed to use of word order so the system is considering semantic of the essays. And LSA is well suited for Science, medicine, Social and business not for all essays.

Table 1. Term document matrix m*n:

|       | DOC1 | DOC2 | DOC3 |
|-------|------|------|------|
| Term1 | 2    | 3    | 1    |
| Term2 | 2    | 2    | 1    |
| Term3 | 3    | 2    | 2    |

Table 2 Word assignment matrix to topics m*n singular matrix

|       | Topic 1 | Topic 2 | Topic 3 |
|-------|---------|---------|---------|
| Term1 | 2       | 2       | 1       |
| Term2 | 2       | 1       | 2       |
| Term3 | 1       | 1       | 3       |

Table 3 Topic distribution across document

|      | DOC1 | DOC2 | DOC3 |
|------|------|------|------|
| Topic1 |     |      |      |
| Topic2 |     |      |      |
proposed is sentence fragmentation that is the range of sentences between 15-20 words each. And the model implemented is lexical-semantic techniques [1] based on small set of corpus. And it works for domain-specific, concept specific and concept grammar. Educational Testing Service system uses Microsoft Natural Language Processing (MsNLP) [4] tool for training the model. Before using the MsNLP [4] tool we should manually remove any suffixes and stop words. After removing the stop words classify the sentences for identify weighted words from the sentences.

For structural scoring, we should manually construct the grammar rules. And a separate feature extractions program is used to embedding the sentences’ with parts of speech and preposition phrases, etc. And new essays are given as input to the phrasal node extraction program. But the main drawback is, the system works on preprocessed essays. And most of the preprocessing essay work is done manually.

2.4 E-Rater

[4]E-Rater scoring system was implemented by Burstein Kukich, etc. They proposed an approach which uses both statistical based features and a natural language processing techniques. The futures it retrieves like a variety of words and syntactic structure. And the system was implemented with five modules to score an essay. Out of 5 modules 3 modules will be used for retrieving features and find relevance scores.

The modules are syntax module, discourse module, topic analysis module and other 2 modules are model building module and scoring module are used for scoring essays.

Syntax module: syntactic analyzer (parser) used to identify syntactic constructions in essay text. It uses syntactic “chunker” to tag each word for parts of speech to find phrases and assembles the phrases into trees based on Sub categorization information for verbs.

Discourse module: After Syntax module the Discourse module will identify the discourse type. There 4 types of discourse annotations
1. Narration
2. Exposition
3. Description
4. Argument

By identifying key words, selected terms, and syntactic structures we can categorize each essay according to selected features. Then based on the selected annotations will classify each and every essay to which it belongs to.

Topical Analysis Module: It is used to identify vocabulary used in the easy by using the vector-space model. First total essays are divided into a training set and testing set. Then we convert training and testing essays into vectors with word frequencies. Now we have a training vector and testing vector then will find a similarity between two vectors.

Model Building: It takes the output from all 3 modules and builds a model. It works on linguistics based features [8] and regression technique is used for e-rater model building. To train models, a training set of human scored sample essays is collected that is representative of the country of scores in the scoring guide.

The model building module is a program that works on regression technique. Syntactic, discourse and topical analysis information for the model building sample are used as input to the regression program. The program outputs the predictive features and their associated regression weightings. This output composes the model that is then used for scoring.

Scoring: A linear regression is used to compute the final essay score. To compute the final score for each essay, the sum of the product of each regression [5] weighting and its associated feature integer is calculated.

2.5 Conceptual rater [6]:
C-Rater is an NLP based system to score the essays. And not required large collection of graded essays only required a single correct answer. It considers features like Synonyms, similar words, spelling,
canonical representation of each response. It first identifies misspelled words when the tokens are not founded.

**SYNTACTIC VARIETY**: At the very first, the syntactic analysis extracts the structure of the tuple of each sentence from the essay. And it makes sure that each tuple will consist of a verb clause along with arguments. After finding the tuples, the syntactic analysis eliminates incorrect responses along with its arguments.

**PRONOUN RESOLUTION**: In the next step, the Conceptual Rater finds the pronoun references in the essay. It also identifies noun components that precede the pronoun references. With that, it is able to know all pronoun components are referring noun components in the entire essay.

**MORPHOLOGY**: The morphological analysis phase in Conceptual Rater recognizes two kinds of morphological variations: inflectional and derivational. Inflectional morphology inspects the grammatical markers that attach to words. Derivational morphology involves a change in the syntactic category of a word.

**FILLING IN THE SEMANTIC GAPS**: Finally, it constructs a word similarity matrix to fill the semantic gaps with using a vector model.

2.6. Bayesian essay test scoring system [3]: BETSY is being developed by Lawrence M. Rudner. It classifies an essay into a four-point nominal scale (e.g., extensive, essential, partial, and unsatisfactory). It classifies the essays on both content and style specific. It uses two models to classify 1. Multivariate Bernoulli Model 2. Bernoulli Model

**Surface features**: Number of words, number of verbs, number of commas, frequency of article, and average sentence length

**Content features**: LSA [9], specific words, frequency of content words, phrases.

BETSY uses probability to classify the essays into 3 categories.

1. The Probability that the features are included in the essay given that the examiner has provided an appropriate response.
2. The Probability that the features are included in the essay given that the examiner has provided a partially appropriate response.
3. The Probability that the features are included in the essay given that the examiner has provided an inappropriate response.

2.7. Intelligent essay marking system:

EMS is developed by Ming, etc all at NGEE ANN Polytechnic and it is based on Pattern Indexing Neural Network. The system can be used both as an assessment tool and for content and structural purpose. It uses an algorithm called Idextron it is a curationization Algorithm that performs pattern recognition and in this case the patterns are the words of the texts. The Idextron-based neural network attempts to overcome a slow, no incremental training, which is typical of traditional Artificial Neural Networks [7][1][10]

2.8. Automark: was developed by Mitchell, etc all. Automark uses NLP techniques to assess all type responses. Automark takes features like Punctuation and spelling. Syntactic constitutes, pattern matching. First, the responses are pre-processed to standardize the input in terms of punctuation and spelling. Then, a sentence analyzer is used to identify the main dependency of text in an essay that how they are related. After that a pattern-matching algorithm is used to searches for similarities between the human rater text (which are marked by expects) and the syntactic constituents of the student text. Finally, the feedback module gives the feedback on student responses related pattern match. But more specific feedback is expected to be possible.

2.9. Schema extract analyze and report: The SEAR is proposed by Christie. It has two different methods to assess the style and content of an essay. For the style assessment, it uses pre-determining candidate metrics, first, it divides the essays into training and testing essays then it takes training essays to marks them manually according to pre-determined metrics. Then after it starts assigning weights randomly for each metric until the error rate is minimized between human rated to computer marking essays. And after finding the weights of each metric. It tests the weights on the tested essays and finally assigns the score for the essays.
2.10. **Paperless school free-text marking engine:** It is a Web-based Learning Management System developed by Mason and Grove-Stephenson. It uses the NLP [2] technique to assess the essay. First, it collects master text from different resources such as textbooks, encyclopedias or relevant websites. And model takes responses from student to evaluate. These student responses are compared with master text to find the score. The grades given by the marker and those resulting from the combination of the parameters; and finally result will be provided to the students. These all processes will be done in three processes called Knowledge level, Understanding level, Evaluation level.

2.11. **Automated text marker:** The Automated Text Marker (ATM) was developed by Callear, etc all at Portsmouth University. And this system also assessing the essays based on content and style based features for these the system is designed to assign two different scores for the given essay one for the content and another for the style. And the teacher should combine the score of two modules to give the final score. For that, it uses NLP [2] [13] and information extraction techniques to assess the student essays.

The system works on two parallel modules the At first is syntax analyzer, takes all essays as input and retrieves the features like number of sentences and then checks for grammar and sentence fluency. And the semantic analyzer looks for concepts in the sentences and their dependencies to each other in domain level. Then a pattern-matching procedure is used to map the text between the student’s responses and the reference model response.

2.12. **Apex assessor:** Apex Assessor is developed by Dessus, Lemaire, and Vernier. It is having web-based learning environment. It provides a user interface where the student can write text and one can self assess the essays after that it provides feedback on the essay based on style and coherence with this the student can improve the writing skills. And it asses the essays on two parts one content-based and another is style based.

- **Content-based assessment module:** in content-based first it converts essays into vector form and represents the essay in LSA format then compares the student representation answer with the trained model.

- **Outline assessment module:** this module provides a similar portion of the essay for each paragraph in the student’s text so that the student can be given an outline view of the essay.

- **Coherence assessment module:** this module finds the semantic distance between student vectors to expert vector. After finding the difference it compares with threshold value if it is below the threshold value it intimate to the student.

2.13. **Auto-marking:** Auto-marking was developed by Pulman, etc all. Its focus is not to automatically score the essays, but to help in low standard students by providing feedback. It asses the essays and gives an integer value like (0,1,2) 0 indicates poor essay lack in style and content, 1 indicates partially correct but their chance to improve the quality of the essay. 2 indicates the essay is well written no need of any modification. For this it uses NLP [2] [16] and, pattern matching techniques.

Customization and shallow processing module: it uses a Hidden Markov Model [5] [2][14] to tag part-of-speech (POS) to each and every word in sentence.

The pattern-matcher module: It uses pattern matching technique to match the student essays with expert essays and gives the similarity score.
2.14. **CarmelTC**: This module has been developed at the University of Pittsburgh by Rosé, Roque, Bhembe and Vanlehn. Apart from giving a score to the student, it will give feedback on the student essays like which part of the essay is correct and which is not correct with help of some classification algorithms like Naive Bayes classification [3][15] and Carmel’s linguistic analysis and with the help of machine learning it gives a score. In the first step it breaks the essay into sentences, Next Bayesian classifier is used to retrieve the features from each sentence. And for each sentence a vector is prepared with retrieved features. And finally identifies sentence classes based on these feature vectors with the ID3 tree learning algorithm.

2.15. **Intelli metric**: It is a commercial system whose focus is mainly on style and content based assessment. First, it requires an initial training phase which is trained on manually scored answers by human graders’. In this phase based on the essay type, it will finalize the important features and gives metric to every each feature. The features may vary from essay to essay. After training the model it will apply the model on another sort of essays to score the essays.

2.16. **Larkey’s system**: Larkey has proposed the system it works on binary text categorization techniques to assess students’ essays by classifying them as “good” or “bad”. It considers both their content and their style[17-19].

**Bayesian classifiers [3]**: First, it assigns probability score for each document based on the previously specified category of documents. For this, it selects some number features by removing stop words, and by applying stemming on each word then it applies Bayesian networks[3] [11][12]to train the binary model. This network classifies the essays into two categories like label values 0 and 1. Bayesian classifies retrieve the features automatically calculated from the text. It takes the features like number of characters in the document, the number of different words in the document, the average sentence length, the number of sentences, the average word length and the number of words longer than seven characters
### Table 4. Comparison of all essay grading systems

| S.NO | System                                      | Features                                                                 | Technique                                      | Training                                                                 |
|------|---------------------------------------------|--------------------------------------------------------------------------|-----------------------------------------------|--------------------------------------------------------------------------|
| 1    | PEG                                         | Sentence structure(essay length, preposition count, POS, grammar, relative pronouns) | Statistical Approach(linear Regression)       | Training based on previously marked essays                               |
| 2    | Intelligent Essay Assessor (IEA)            | Bag of words                                                             | Topic modeling, Latent Semantic Analysis, Singular Value Decomposition | Student essays and class teacher essay                                    |
| 3    | Educational Testing Service (ETS I)         | Remove suffixes and stop words(manual), Metonyms(manual classification)  | MsNLP tool, Phrasal node extraction           | Domain specific, Concept specification                                   |
| 4    | Electronic Essay Rater (E-Rater)            | Variety words used syntactic structure(POS), Syntax, discourse topic, Topic analysis (vocabulary usage, vector space model, weight vector) | MsNLP tool, Corpus based approach, Corpora linguistic approach | Trained with set of essays scored by faculty                               |
| 5    | Conceptual Rater (C-Rater)                  | Synonyms, similar words, spelling, canonical representation of each response | Distance algorithm, pronoun resolution, filling semantic gaps | No large Collection of graded essays                                      |
| 6    | Bayesian Essay Test Scoring System (BETSY)  | Surface features(word count, verbs count, commas count, sentence length, frequency of article), Content features (specific words, frequency of content words, occurrence of specific noun verbs) | Multivariate Bernoulli model, Bernoulli model, probability |                                                                       |
| 7    | Intelligent Essay Marking Systems (IEMS)    | Pattern recognition                                                      | Pattern indexing neural networks, Indexon clusterisation Algorithm |                                                                       |
| 8    | Automark                                    | Punctuation and spelling, Syntactic constitutes, pattern matching         | NLP                                           | NLP                                                                      |
| 9    | Schema Extract Analyse and Report (SEAR)    | set of common metrics, some initial calibration                          | NLP                                           | subset of essays as training                                             |
| 10   | Paperless School free-text Marking Engine (PS-ME) | correct master texts for comparison                                      | NLP techniques                                |                                                                         |
| 11   | Automated Text Marker (ATM).                | text, and their dependencies, syntax, semantic analysis                  | IE techniques, NLP                            |                                                                         |
| 12   | Apex Assessor                               | most similar portion, semantic distance between sentences                | Latent Semantic Analysis, NLP, set of unmarked texts for training |                                                                         |
| 13   | Auto-marking                                | part-of-speech tagger                                                    | NLP and pattern-matching techniques           |                                                                         |
| 14   | CarmelTC                                    | text in sentences                                                        | Carmel’s linguistic analysis, 126 physics essays |                                                                         |
3. Observations

All 16 essay grading systems are concentrating on style and sentence arrangement on statistical features using some machine learning models and some systems are working on content-based essay scoring adding their not used fully NLP techniques so our feature work is on fully content-based scoring and finding essay relevance score using deep learning and natural language processing techniques.

4. References

[1] Valenti, Salvatore, Francesca Neri, and Alessandro Cucchiarelli (2003). "An overview of current research on automated essay grading.” Journal of Information Technology Education 2, 319-330.
[2] Shristi Drolia, Shrey Rupani, Pooja Agarwal and Abheejeet Singh. (2017) “Automated essay rater using natural language processing,” International Journal of Computer Applications 163(10)
[3] Rudner, Lawrence M., and Tahung Liang (2002). Automated essay scoring using bayes' theorem. The Journal of Technology, Learning and Assessment 1(2).
[4] Powers and Donald E. 2001 "Stumping E Rater: Challenging the validity of automated essay scoring.” ETS Research Report Series 1(1).
[5] Dikli and Semire. 2006 “An overview of automated scoring of essays.” The Journal of Technology, Learning and Assessment 5(1).
[6] claudia leacock1 and martin chodorow, (2003) C-rater: Automated scoring of short-answer Questions Computers and the Humanities 37: 389–405.
[7] Burstein, J., Chodorow, M., and Leacock, C. (2004). Automated Essay Evaluation: The Criterion Online Writing Service. AI Magazine, 25(3), 27
[8] Hengbin Yan and Jonathan Webster, a Corpus-Based Approach to Linguistic Function, PAACLIC-27.
[9] WILD, 2005. Parameters driving effectiveness of automated essay scoring with LSA. IN: roceedings of the 9th CAA Conference, Lough- borough: Loughborough University
[10] Ramalingam, V. V.; Pandian, A.; Chetry, P. and Nigam, H.(2018) Automated Essay Grading using Machine Learning Algorithm Journal of Physics: Conference Series, IOP Publishing, 1000
[11] harshavardhan, a. (2019)“techniques used for clustering data and integrating cluster analysis within mathematical programming.” journal of mechanics of continua and mathematical sciences,14(6),
[12] s. n. p. and al. (2019) “variation analysis of artificial intelligence, machine learning and advantages of deep architectures”, ijast, 28(17) 488 - 495
[13] sallauddin md (2019)“a comprehensive study on traditional ai and ann architecture.” International Journal of Advanced Science and Technology,28(17) 479–87
[14] D. Kothandaraman, M. Shesikala, K. SeenaNaik, Y. Chanti and B. Vijayakumar.(2019) “Design of an Optimized Multicast Routing Algorithm for Internet of Things”, International Journal of Recent Technology and Engineering, 8(2).
[15] Mahender K, Kumar TA and Ramesh KS 2018 An efficient FBMC based modulation for future wireless communications ARPN Journal of Engineering and Applied Sciences 13(24) 9526-9531
[16] Jamalpur Bhavana, Komuravelly Sudheer Kumar( 2019). Implementation of Bovw Model Towards Obtaining Discriminative Features of the Images, International Journal of Advanced Science and Technology, 28(17).
[17] Rajasri I, Guthpa AVSSKS and Rao YVD 2011 Influence of Structural Aspects on the Generation Process in Planetary Gear Trains Engineering 3(10) 1018-1021 DOI: 10.4236/eng.2011.310126
[18] P. Pramod Kumar and Dr. K. Sagar, (2019) “A proficient and smart electricity billing management system”, *International Conference on Emerging Trends in Engineering and published in Springer Nature as a Part of the Learning and Analytics in Intelligent System 3*

[19] Seena Naik K and Sudarshan E 2019 Smart healthcare monitoring system using raspberry Pi on IoT platform *ARPN Journal of Engineering and Applied Sciences* 14(4) 872-876.