International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

An Analysis of Automated Essay Grading Systems

Kshitiz Srivastava, Namrata Dhanda , Anurag Shrivastava

***Abstract**: Essays are one of the most important method for assessing learning and intelligence of a student. Manual essay grading is a time consuming process for the evaluator, a solution to such problem is to make evaluation through computers. Many systems were proposed over past few decades. Each system works on different approach having focus on different attributes. Aim of this paper is to understand and analyze current essay grading systems and compare them primarily focusing on technique used, performance and focused attributes.*

***Keywords**: Automated Essay Grading, Computer-based Assessment Systems, Text Processing, Essay Evaluation, and Semantic Analysis.*

I. INTRODUCTION

Essays are considered as one of the main evaluation criteria used by teachers to evaluate student's performance. Essay evaluation is a time consuming process, a teacher denotes a huge amount of time in evaluation of essays because of its subjectivity. Also because of subjective nature of the essay variation in grades usually occurs. Solution to such problem is automatic essay evaluation. Evaluating essays through computer will help reducing teachers load as well as reduce the variation in grades as a result of human factors. Many system were developed/ proposed to check the writing quality of the essays some of the mentions are Project Essay Grade (PEG), Intelligent Essay Assessor (IEA), Educational Testing service I, Electronic Essay Rater (E-Rater), C-Rater, BETSY, Intelligent Essay Marking System, SEAR, Paperless School free text Marking Engine and Automark etc. Most of them are either commercially available or under development. This paper aims on the reviewing and comparing above mentioned essay grading system.

II. RELATED WORK

Over the past decade focus of automation of essays becomes an important challenge for educators. With the increase in e-learning market demand for automation of evaluation process becomes the need of the hour.

Manuscript received on March 15, 2020.
Revised Manuscript received on March 24, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Kshitiz Srivastava*, Scholar, Ph.D. scholar, Dr. APJ Abdul Kalaam technical university, Lucknow, Uttar Pradesh, India
Prof. (Dr.) Namrata Dhanda , Professor, Department of computer science, Amity University, Lucknow, Uttar Pradesh, India
Dr. Anurag Shrivastava , Professor, Department of Computer Science & Engineering, BabuBanarasi Das Northern India Institute of Technology, Lucknow, Uttar Pradesh, India

We have many online evaluation systems that are based on objective answers, only a hand full of systems are available that can actually evaluate the subjectivity of answers. This paper is based upon various papers published on different automatic essay evaluation systems. Due to proprietary and commercialization of many systems, much information was not available.

The first system developed was **Project essay grader (PEG)** is one of the earliest developed system which was able to automatically evaluate the essay. It was developed by Hearest and Page[1] in the mid on 1960's and is probably the first automatic assessment tool, Page claims that computer can grade essay better than humans with the help of project essay grader [12]. Initially its accuracy was less but later features were added and more accuracy was gained. Project essay grader uses set of pre-graded essays and extract the linguistic features from it. After feature extraction multiple linear regressions is applied to determine optimal weighted features [4]. Drawback of project essay grader is that it focuses on the writing style rather than content. As initially developed it uses only statistical approach. Pages latest experiment shows regression correlation of 0.87% with human grading [1].

One of the major drawbacks of PEG was its no focus on the text semantics. To overcome this issue Hearest and team develops a new system, **intelligent essay assessor** which is based on latent semantic analysis [2]. It computes and combines content, style, and mechanics of the essay[5]. Latent semantic analysis considers the semantic space of each term in high dimensional semantic space. In latent semantic analysis firstly we compute two dimensional semantic spaces where columns and row represents document and frequency of words respectively. Then this matrix is decomposition using singular value decomposition. After singular value decomposition cosine similarity is used to measure the similarity of the answer with the pregraded essay [1].

One of the major drawback of latent semantic analysis is that it does not take word order into account tough it was not considers as an important feature [2] in IEA. Performance of Intelligent essay assessor is 85%-91% accuracy in comparison with human grader, based on the test conducted on GMAT essay [1].

Around same time span a different system was being develop, **Educational Testing Service (ETS I),** it was developed by Burstein. This system works on small sentences. From the training data a domain specific and concept based lexicon is developed. Generation of lexicon requires suffix and stop words removal. For parsing Microsoft natural language processing (MsNLP) Tool[1] is used. Set of linguistic features that might more directly measure could automatically extract from essays using NLP and IR techniques. The team had found more than hundred extract-able features from the essay that can be later used in essay grading [5].

Test programs like SAT, GRE, TOEFL etc. uses ETS I [9]. It has accuracy of 80-90% depending whether essay marking or marking both essay and test set [1]. Burstein also developed another system which was based on natural language processing and statistical technique called **Electronic Essay Rater (E-Rater)**, it uses Microsoft natural language processing tool for parsing sentences. A set of pre-graded essay are required to create a standard reference for other answers [1]. E-Rater firstly extracts Vocabulary content, syntactical information etc. and with the help of multiple linear regression essay is graded. E-rater have five modules, out of five, three modules identify features and the vocabulary usage of an essay remaining two modules are used to select and weigh predictive features for essay scoring and scoring [3]. Its accuracy lies between 87-94% [1]. Major drawback of previous discussed systems is the requirement of a large number of pre-graded essays to build training set and the focus on statistical attribute rather than concept of the essay. To overcome this problem **Conceptual Rater (C-Rater)** was developed. It is based on natural language processing. It judges the answer for the correctness of the essay [1]. As stated it does not requires large number of previously graded essay, instead a single evaluated answer can work fine. C-rater evaluates analytic based content. C-rater is not so popular, rater it is used with other grading tools. Its accuracy is about 80% [1]. Another model based on features of Project essay grader, E-rater and latent semantic analysis was developed by Lawrence M. Rudner called **Bayesian Essay Test Scoring system (BETSY).** It is a window based program written in power basic, BETSY is based on Bayesian models. There are 2 models multivariate Bernoulli model and the multinomial model. In multivariate Bernoulli model the probability of presence of a feature is estimated by the proportion of essays within each category that include the feature whereas In multinomial model, on the other hand, the probability of each score for a given essay is computed as the product of the probabilities of the features included in the essay. It classifies essay into a four point nominal scale (e.g. extensive, essential, partial, unsatisfactory) using a large set of features including both content and style specific issues [11]. Accuracy of 80% with described data set [1]. Another system, **Intelligent Essay Marking Systems (IEMS)** was developed at NGEE ANN Polytechnic. It is based on pattern indexing neural network and uses a specialized clustering algorithm called "Indextron" [1]. This system is capable of quick feedback, which was not available with other systems [3]. This system is use to grade qualitative essay and maintained correlation of 0.8. With the enhancement of essay grading systems, need for evaluating free text answers for open ended questions also increased. **Automark** developed to solve this problem. Created by Mitchel and team in late nineties, Automark focuses on free text answers and open ended text. It uses information extraction and some natural language processing technique for response grading. It looks for specific content in the free text answers ignoring spelling, typing or semantic errors [1]. Various formats are specifies in which answer can be given each template represents one form of a valid or a specifically invalid answer. Automark consist of following steps [11]. Its correlation lies between 93- 96%. Essays are evaluated on the basis of three attributes namely writing style, content and semantics. A system **Schema Extract Analyze and Report (SEAR)** was developed by Christie as a final year project in Robert Gordon University, Aberdeen. It provides method of evaluation for both essay content and essay style. Methodology for evaluation based on style uses set of common metric. For content based evaluation neither training nor calibration is required, teacher creates references for assessment. It uses information extraction techniques to fill the student's schemes with the student's data and to compare them against the references. The content schema is prepared once and revised regularly [11]. In 2002 Paperless School free-text Marking Engine (PS-ME) was presented by Mason and Grove Stephenson in the Birmingham University in 2002. It is used for assessment for both summative and formative answers. It is mainly used in web-based learning. Because it's slow processing it cannot grade in real-time. NLP is the core of PS-ME for assessment. A concept of master text is used, it is a reference schema with which essay is compared and graded. It may also contain negative master text that contains the common mistakes done by student for that particular answer. Essay is compared with every master text; this evaluation is calculated through linguistic analysis. The weights are derived during the initial training phase.

In year 2016 Alikaniotis and team introduced model based on deep neural network which was able to learn features automatically. They introduced a new method which can identify the regions of the text with the help of Score-Specific Word Embedding (SSWE) and a two-layer Bidirectional Long-Short-Term Memory (LSTM)[23]. They have extended the C&W Embedding model to capture local linguistic features of each word as well as how each word is used in total score of an essay. They used the Kaggle's contest dataset having 12.976 essays each double marked (Cohen's = 0.86). The essays presented eight different prompts, each with distinct marking criteria and score range.

Dasgupta and team, in year 2018 proposed a new "Qualitatively enhanced Deep Convolution Recurrent Neural Network architecture". The model focuses on word-level as well as sentence-level representations of essay for evaluation. They consider many features in the text [24]. Their architecture for the Convolution RNN has five layers: Generating Embedding Layer, Convolution Layer, Long Short-Term Memory Layer, Activation layer and Sigmoid Activation Function Layer. They also used Kaggle's ASAP contest dataset.

There are three major attributes around which the above system works. These attribute are writing style, content and semantics. The style attributes deals with the lexical sophistication, mechanical and grammatical aspect of the essay. Second types of attributes are content attribute which are mainly based on comparing the students essay with a pre-evaluated one. The third attribute is semantic attribute; semantic attribute deals with the meaning behind the text. Most of the systems discussed any deal in any of the combination of the three given times of attribute. A systematic comparison on the technique used, performance, application and the attribute used is done in table 1.

III. RESULT ANALYSIS

Automated Essay grading has started in mid 1960 with project Essay grader developed by Hearest and Page with the focus on style attribute of the essay, since it was developed in mid on 1960 it uses only statistical approaches only having correlation of 0.87 with human grader. To overcome drawbacks of PEG, Hearest and team develop intelligent essay assessor, it focuses on content attribute of the essay. This system was based on latent semantic analysis with aggregation of 0.85%. Parallel to IEA, Burstein and team developed Educational Testing Service (ETS I) which was probably first system that works upon Natural language processing and Information retrieval. It also evaluate essay for content attribute with accuracy between 93 to 96% based on different experiment. Burstein also developed a system known as E-RATER which uses a blend of statistical approaches and natural language processing. It focuses on style and content attribute both with aggregation of 0.85%. Another system C-rater was developed based on Natural Language processing, can grade for style and content and have aggregation of 0.8%. Following the track of E-Rater and C-Rater many tools like BETSY, IEMS, Automark and SEAR can grade for both style and content both using various techniques like statistical approaches, natural language processing, rule based expert system etc.

SAGrader, based on rule based expert system can grade essay for shallow semantic features and is still under process. Similarly SAGE, based on Natural language processing can grade essay for semantic attribute and is still under development.

In recent years automated essay grading systems are focusing on style and content as proposed by Alikaniotis and Dasgupta.

System	Technique	Performance	Application	Attributes
PEG	Statistical	Corr:0.87	Nonfactual disciplines	Style
IEA	LSA	Agr:0.85	Psychology and military essays	Content
ETS I	NLP	Acc: 93-96%	GRE, SAT, CLEP etc.	Content
E-RATER	Statistical/NLP	Agr:0.87	GMAT exam and English writing	Style and Content
C-RATER	NLP	Agr:0.8	Reading comprehension and algebra	Style and Content
BETSY	BETSY/Statistical	Acc: 80	Not Available	Style and Content
IEMS	Indextron	Corr:0.8	Non-mathematical Essays	Style and Content
AUTOMARK	IE	Corr:0.93	Statutory NCA of science	Style and Content
SEAR	IE	Corr:0.3	History essays	Style and Content
PS-ME	NLP	NA	NCA or GCSE exam	Style
SAGrader	rule-based expert systems	NA	Under process	Semantics
SAGE	NLP	NA	Under process	Semantics
Alikaniotis, Yannakoudakis&Rei (2016)	SSWE + Two-layer Bi-LSTM	~0.91 (Spearman) ~0.96 (Pearson)	Under process	Style and Content
Dasgupta et al. (2018)	Deep Convolution Recurrent Neural Network	Pearson's 0.94 Spearman's 0.97	Under process	Style and Content

Table 1: Comparison between various AEE System

IV. CONCLUSION

Over the past few decades, many approaches and systems on scoring essay questions has been developed. These Automated essay scoring systems basically find features in the essay and grade for them. We can divide these features in terms of three attributes namely: style, content and semantic.

Most of the system proposed works on either style, content or both and only two SAGrade and SAGE focuses on shallow semantic features, and are still under development. In future we require systems that primarily focus on semantic attributes for essay grading we can also have new semantic attributes that can help in evaluating essay more accurately. Current grading systems cannot detect correctness of the essay we can also have different ways to check the consistency of the essay.

REFERENCES

1. Valenti, S., Neri, F., &Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, *2*(1), 319-330.
2. Williams, R. (2001). Automated essay grading: An evaluation of four conceptual models. In *New horizons in university teaching and learning: Responding to change* (pp. 173-184). Centre for Educational Advancement, Curtin University.
3. Mahato, S. (2017). LEXICO-SEMANTIC ANALYSIS OF ESSAYS IN HINDI.
4. Hearst, M. A. (2000). The debate on automated essay grading. *IEEE Intelligent Systems and their Applications*, *15*(5), 22-37.
5. Hearst, M. A. (2000). The debate on automated essay grading. *IEEE Intelligent Systems and their Applications*, *15*(5), 22-37.
6. Dhokrat, A., &Mahender, C. N. (2012). Automated Answering for Subjective Examination. *International Journal of Computer Applications*, *56*(14).
7. Devi, M. S., & Mittal, H. (2016). Machine learning techniques with ontology for subjective answer evaluation. *arXiv preprint arXiv:1605.02442*.
8. Guruji, A., Mrunal, M. P., Pawar, S. M., &Kulkarni, P. J. (2015). Evaluation of Subjective Answers Using GLSA Enhanced with Contextual Synonymy. *Int. J. Natural Language Computing*, *4*(1).
9. https://www.ets.org (ETS official website)
10. Sukkarieh, J. Z., & Blackmore, J. (2009, March). c-rater: Automatic content scoring for short constructed responses. In *Twenty-Second International FLAIRS Conference*.
11. Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1).
12. Abbas, A. R., & Al-qaza, A. S. (2014). Automated Arabic Essay Scoring (AAES) using Vector Space Model (VSM). *Journal Of AL-Turath University College*, (15), 25-39.
13. Page, E. B. (1967). Statistical and linguistic strategies in the computer grading of essays. In *COLING 1967 Volume 1: Conference Internationale Sur Le TraitementAutomatique Des Langues*.
14. Attali, Y. (2007). ON-THE-FLY CUSTOMIZATION OF AUTOMATED ESSAY SCORING. *ETS Research Report Series*, *2007*(2), i-25.
15. Zupanc, K., &Bosnic, Z. (2014, December). Automated essay evaluation augmented with semantic coherence measures. In *2014 IEEE International Conference on Data Mining* (pp. 1133-1138). IEEE.
16. Zupanc, K., &Bosnic, Z. (2016). Advances in the field of automated essay evaluation. *Informatica*, *39*(4).
17. Darwish, S. M., & Mohamed, S. K. (2019, March). Automated Essay Evaluation Based on Fusion of Fuzzy Ontology and Latent Semantic Analysis. In *International Conference on Advanced Machine Learning Technologies and Applications* (pp. 566-575). Springer, Cham.
18. Azmi, A. M., Al-Jouie, M. F., &Hussain, M. (2019). AAEE–Automated evaluation of students' essays in Arabic language. *Information Processing & Management*, *56*(5), 1736-1752.
19. Mehmood, A., On, B. W., Lee, I., & Choi, G. (2017). Prognosis essay scoring and article relevancy using multi-text features and machine learning. *Symmetry*, *9*(1), 11.
20. Ajetunmobi, S. A., &Daramola, O. (2017, October). Ontology-based information extraction for subject-focussed automatic essay evaluation. In *2017 International Conference on Computing Networking and Informatics (ICCNI)* (pp. 1-6). IEEE.
21. Ke, Z., & Ng, V. (2019, August). Automated essay scoring: a survey of the state of the art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (pp. 6300-6308). AAAI Press.
22. Goh, T. T., Sun, H., & Yang, B. (2019). Microfeatures influencing writing quality: the case of Chinese students' SAT essays. Computer Assisted Language Learning, 1-27.
23. Alikaniotis, D., Yannakoudakis, H., &Rei, M. (2016). Automatic text scoring using neural networks. arXiv preprint arXiv:1606.04289.
24. Dasgupta, T., Naskar, A., Dey, L., &Saha, R. (2018, July). Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications (pp. 93-102).
25. Ke, Z., & Ng, V. (2019, August). Automated essay scoring: a survey of the state of the art. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (pp. 6300-6308). AAAI Press.
26. Nadeem, F., Nguyen, H., Liu, Y., &Ostendorf, M. (2019, August). Automated Essay Scoring with Discourse-Aware Neural Models. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 484-493).

AUTHORS PROFILE

Mr. Kshitiz Srivastava is currently pursuing his Ph.D. from Dr. A.P.J. Abdul Kalaam Technical University, Lucknow, Uttar Pradesh, India in the field of Computer Science and engineering. He completed his M.Tech in the field of Computer Science and engineering. His area of interest is Programing Languages, Theory of automata and formal languages, and Data Mining.

Prof (Dr.) Namrata Dhanda is currently working as Professor in the Department of Computer Science, Amity University, Lucknow, Uttar Pradesh, India. She has 20 years of experience in teaching. She had completed her Ph.D. at University of Petroleum and Energy Studies, Dehradun Uttarakhand, India. Her area of research includes semantic web, machine learning, Activity Theory. Her subjects of interest are Theory of automata and formal languages, Database Management Systems, Design and Analysis of Algorithms, Compiler Construction to name a few. She has published papers in several National and International Conferences and journals of repute.

Dr. Anurag Shrivastava is currently working as Professor in the Department of Computer Science & Engineering, Babu Banarasi Das Northern India Institute of Technology, Lucknow, Uttar Pradesh, India. He has more than 18 years of experience in teaching. He had completed his Ph.D. at Dr. A.P.J. Abdul Kalaam Technical University, Lucknow, Uttar Pradesh, India. His area of research is requirements engineering process assessment and improvement. His area of interest is design and analysis of algorithms, Theory of automata and formal languages. He has published SCI and scopus indexed journals.