International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 124-127
Published Online March 2021 in IJEAST (http://www.ijeast.com)

SURVEY PAPER ON SMART ESSAY GRADER

Prof.Nidhi Sharma, Aman Sharma, Vaibhav Urkude, Adhoksh Sonawane
Department of Computer
Bharati Vidyapeeth College of Engineering, Navi Mumbai, Maharashtra, India

Abstract—**In order to optimize Human-Machine agreement for automatic evaluation of textual summaries or essays, automated essay grading has been a research field. With a growing number of people taking multiple exams such as the GRE, TOEFL, and IELTS, grading each paper would become more challenging, not to mention the challenge for humans to maintain a consistent mindset. In this situation, it is extremely difficult to rate a large number of essays in a short amount of time. This project aims to address this issue by developing a stable interface that will aid humans in grading essays. This study served as a medium for us to extract features such as the Bag of Words, numerical features such as the count of sentences and words, as well as their average lengths, structure, and organization, in order to rate the essay with the highest level of accuracy. This algorithm was chosen because it works well for small datasets.**

Key Words: **NLP, Essay, Words, Sentences, NN.**

1. INTRODUCTION

1.1 Purpose

Since the early 1960s, automated essay grading has been a research subject. It's a challenging job because we need to extract both quantifiable and nonquantifiable attributes, such as the writer's feelings, when writing on paper. It will appear that extracting is a simple procedure. The system's goal is to divide a large number of textual entities into a small number of distinct groups, each corresponding to a range of possible scores—for example, 1-100. The artificial environment we generate will recognize patterns and try to predict the next possible performance using a training dataset. This project investigates how text mining can aid in essay scoring.

The method of grading student essays without human intervention is referred to as automated essay scoring. An AES system takes an essay written for a specific prompt as input and assigns a numeric score to the essay based on its material, grammar, and organization. Regression methods are normally applied to a collection of carefully constructed features in such AES systems. Humans find it difficult to understand all of the considerations that go into awarding a grade to an essay. Since the system is focused on recurrent neural networks, it can effectively encode the information needed for essay evaluation while also learning complex patterns in the data through non-linear neural layering. The results show that the system outperforms a strong baseline in automated essay scoring and achieves state-of-the-art success.

1.2 Product Scope

The aim is to translate the method into one of the human languages, with all of the complexities that entails. We've even looked for algorithms to see if they're correct. Furthermore, this has allowed us to learn more about automated systems and test them with a machine learning algorithm to create a stable interface that will serve our needs.

2. LITERATURE SURVEY

We have gone through a number of papers related to our project which had been pursued by number of researchers in the past. From these papers, we got to know the different technologies which were used by different researchers to implement this project. Thus, we have prepared a literature survey on some of most clean and efficient systems which were used by researchers to implement the project.

Systems mentioned in technical paper	Publisher	Approach	Correlation with human scorers
1. Automated Essay Scoring	Dong and Zhang	CNN	~0.7344
2. BETSY	Rudner	Bayesian text classification	~0.80
3. Project Essay Grader(PEG)	Ellis Page	Statistical	0.87

4. E-rater	ETS development team	NLP	~0.90
5. Intelligent Essay Assessor(IEA)	Landauer, Foltz, & Laham	LSA (KAT engine by Pearson)	0.90

Table 1: Table of past systems of Essay graders

1. Intelligent Essay Marking Systems (IEMS)

IEMS is based on the Pattern Indexing Neural Network (the Index Tron) developed at NGEE ANN Polytechnic (Ming, Mikhailov, & Kuan, 2000In several content-based subjects, the framework can be used as an evaluation tool as well as for diagnostic and tutoring purposes. Students may receive immediate feedback and discover where and why they performed well or poorly. As a result, it can be integrated into an intelligent tutoring framework that will assist students in improving their writing skills by easily grading papers and providing input. For IEMS, the normal protocol has been to begin with a practice collection of essays that have been meticulously hand-scored. The software assesses surface features of each essay's text, such as the total number of words, the number of subordinate clauses, and the ratio of uppercase to lowercase letters-quantities that can be calculated without human interference. After that, it creates a mathematical model that relates these quantities to the scores that the essays received.

PERFORMANCE: According to Ming et al. (2000), an experiment involving the assessment of essays written by 85 students participating in a project report writing module from six third-year Mechanical Engineering classes yielded a correlation of 0.8.

2. Bayesian Essay Test Scoring system (BETSY)

BETSY is a software created by Lawrence M. Rudner at the University of Maryland's College Park with funds from the US Department of Education to classify text based on qualified content.

According to Rudner and Liang (2002) the goal of the system is to determine the most likely classification of an 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. The underlying models for text classification adopted are the Multivariate Bernoulli Model (MBM) and the Bernoulli Model (BM). This method of machine grading combines the best features of PEG, LSA, and e-rater, as well as a few main advantages of its own. It is simple to use, can be extended to a wide variety of content fields, can be used to produce diagnostic outcomes, can be adapted to produce classifications on different abilities, and is simple to communicate to non-statisticians.

PERFORMANCE: Rudner and Liang (2002) report about two text classification models that were calibrated using 462 essays with two score points. The calibrated systems were then applied to 80 new prescored essays, with 40 essays in each score group. An accuracy of over 80% was achieved with the described dataset.

3. Project Essay Grading

PEG is one of the earliest and longest-lived implementations of automated essay grading. It was created by Page and others (Hearst, 2000; Page, 1994, 1996) and is based on style analysis of a block of text's surface linguistic features. Thus, an essay is predominantly graded on the basis of writing quality, taking no account of content. PEG solely relies on a statistical method focused on the premise that observable proxies represent the content of essays. No Natural Language Processing (NLP) technique is used and lexical content is not taken in account. PEG also requires training, in the form of assessing a number of previously manually marked essays for proxies, in order to evaluate the regression coefficients, which in turn enables the marking of new article/composition. The first of the automated essay scorers was Project Essay Grade (PEG). The aim is to go through the history of automated essay grading, why it was impractical when it was first developed, what re-energized growth and research in automated essay scoring, how PEG functions, and what recent research involving PEG has disclosed.

PERFORMANCE: Page's most recent studies with human graders yielded findings with a multiple regression correlation as high as 0.87.

4. Intelligent Essay Assessor (IEA)

The Latent Semantic Analysis (LSA) technique was initially designed for indexing documents and text retrieval, and IEA was developed in the late 1990s. The writers of LSA believe that word order isn't the most important element in grasping the context of a passage, so they don't use it. It also necessitates a large amount of data to create a suitable matrix representation of word use/occurrence, and computations are time-consuming due to the size of the matrices involved. The IEA has a low unit cost, fast personalized reviews, and plagiarism detection as key features. Furthermore, the authors claim that the system is very well suited to analyze and score expository essays on topics such as science, social studies, history, medicine or business, but not suitable to assess factual knowledge. IEA is a valuable domain-independent method that automatically assesses and critiques electronically submitted text essays. It provides immediate input on the student's writing's content and quality. The ability to communicate knowledge orally is a valuable educational accomplishment in and of itself, and one that is undervalued by other types of assessments. Furthermore, essay-based research is thought to facilitate a deeper, more useful level of knowledge and application by students by promoting a stronger conceptual understanding of the subject.

PERFORMANCE: A test conducted on GMAT essays using the IEA system resulted in percentages for adjacent agreement with human graders between 85%-91%.

5. Conceptual Rater (C-Rater)

C-rater is an NLP-based prototype for evaluating short answers to content-based questions, such as those found in the chapter review portion of a textbook (Burstein et al., 2001). C-rater adopts many of some natural language processing tools and techniques developed for E-Rater, even if the two systems differ in many important ways.

For preparation, C-rater does not require a large number of graded responses. Since it is believed impractical to require comprehensive data collection for the purpose of grading relatively low stakes quizzes, it instead uses the single correct answer found in an instructor's guide or answer key, particularly given that a set of short questions is often provided at the end of chapters in a textbook, it uses the single correct answer found in an instructor's guide or answer key. E-rater, a device being built to assess test takers' responses to different types of essay tasks and prompts, was the automated scoring tool used in this study. In a nutshell, C-rater simulates the output of human evaluators using natural language processing techniques. Assessors choose a sample of essays for each essay prompt that have already been scored by at least two human readers and this represents the entire spectrum of potential the end result.

PERFORMANCE: C-Rater achieved over 80% agreement with the score assigned by an instruction.

6. Electronic Essay Rater (E-Rater)

E-Rater extracts linguistic features from the essays to be graded using a mixture of statistical and NLP techniques. Essays are compared to a collection of human-graded essays as a benchmark. An essay that stays on subject, has a solid, coherent, and well-organized argument structure, and uses a variety of words and syntactic structure will obtain a score on the higher end of a six-point scale from E-Rater.

A further feedback component with advisory features has been added to the system. The advisories are based on statistical tests and are entirely separate from the E-Rater ratings, offering additional input on subject and fluency-related aspects of writing. E-Rater was developed using a collection of 270 essays that were manually graded by qualified human raters. Many other available systems are much more complicated and need more preparation than E-Rater. We have an online scoring system that instructors and qualified readers can use to rate text-processed essay responses in addition to handling reader recruiting, training, and observe. When the essay files and reader scores have been scored, they can be submitted to the scoring engine for automated model construction.

PERFORMANCE: Over 750000 GMAT essays have been graded, with human expert and system agreement rates consistently exceeding 97 percent. The empirical findings vary from 87 percent to 94 percent when contrasting human and E-Rater grades across 15 test questions.

3. METHODOLOGY

The Correlation Score is calculated by taking into account all of the characteristics of an essay. Only those features that trigger changes in the score should be considered when processing data sets. Basically, the substance of the essay or the data that the user would enter into the user interface. After saving and sending the essay, it will be saved in text format in the database and sent to a text analyzer for normalization, function abstraction, and data optimization. After the text has been read, it is sent to the NLP Module, which is connected to the Data Analysis Server, which contains all of the tools. And the NLP resources evaluate the text and begin processing the data in the Module, returning a result in the form of a Score and Rubik grade, as well as the number of mistakes. Accuracy is observed by seeing the mean squared value of different NLP models.

We also intend to use the LSTM, a form of recurrent neural network that can learn order dependence in sequence prediction problems. This is a requirement in a variety of dynamic problem domains, including machine translation, speech recognition, and others. Deep learning's LSTMs are a complicated subject. The consistency of this project will then be calculated using the coherent kappa score.

Users explore a broad dataset in an unstructured manner to discover initial trends, features, and points of interest in data discovery, which is the first step in data analysis. Word2Vec is a word embedding learning algorithm that can be applied to large datasets. The conversion of text to vector format is required in order for the NLP module to read and process data. Feature selection is a process in which we try to fit a particular machine learning algorithm into a given dataset. It uses a greedy search method, evaluating all possible feature combinations against the evaluation criteria.

4. CONCLUSION

The use of various methods to study automated essay grading has been done a variety of times. This approach aims to model the language with the most useful features. The outcomes that can be achieved will be both inspiring and valid. We should be able to obtain an average absolute error that is slightly lower than the human standard deviation. As a result, we decided to express through our project that it is possible to have an essay graded automatically, reducing the burden on the individual. All essays will be judged on the same criteria and will receive an accurate score. We'll do whatever we can to keep the best one up to date.

5. FUTURE SCOPE

- There is definitely room for improvement, particularly if we can identify the right features.
- Include deep learning technologies and complex deep learning algorithms such as LSTM and RNN.
- Image processing can also be used to grade handwritten essays that have been graded offline.

6. REFERENCES

- A Neural Approach to Automated Essay Scoring Kaveh Taghipour and Hwee Tou Ng Department of Computer Science National University of Singapore
- AUTOMATED ESSAY GRADING USING FEATURES SELECTION Y.Harika, Sri Latha, V.Lohith Sai, P.Sai Krishna , M.Suneetha(International Research Journal of Engineering and Technology (IRJET))
- PDF, Automated Essay Grading using Machine Learning Algorithm
- Valenti, S., Neri, F and Cucchiarelli, A. 2003 An Overview of Current Research on Automated Essay Grading
- Attali Y and Burstein, J 2006 Automated essay scoring with e-rater
- Automated essay evaluation with semantic analysis Kaza zupac, Zoran Bosnic.
- "Project Essay Grade: PEG", p. 43. In Shermis, Mark D., and Jill Burstein, eds., Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum Associates, Mahwah, New Jersey,
- Larkey, Leah S., and W. Bruce Croft (2003). "A Text Categorization Approach to Automated Essay Grading", p. 55. In Shermis, Mark D., and Jill Burstein, eds. Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum.
- "Neural Net or Neural Network - Gartner IT Glossary". www.gartner.com
- Russell, Ingrid. "Neural Networks Module". Archived from the original on 29 May 2014. Retrieved 2012.
- Graves, A.; Liwicki, M.; Fernandez, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J. (2009). "A Novel Connectionist System for Improved Unconstrained Handwriting Recognition". IEEE Transactions on Pattern Analysis and Machine Intelligence. 31 (5): 855–868. CiteSeerX 10.1.1.139.4502
- Li, Xiangang; Wu, Xihong (2014-10-15). "Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition". arXiv:1410.4281.