Accuracy improvement of short and long answer grading systems using machine learning

Simran Agrawal, Avinash J. Agrawal

Abstract: ***Grading of student answers with the help of language processing techniques has been a defacto standard for automatic marking systems. These systems generally do not take into consideration the errors which might have been introduced by the previous grading systems in order to incrementally improve the grading performance of the system itself. In this paper, we propose a machine learning based algorithm which uses Q-Learning and synset based language processing in order to incrementally improve the automatic grading accuracy for the both short and long answer texts. Usually systems have higher accuracy for short answer matching, but when the same system is applied to long answers then the accuracy reduces drastically. But the proposed algorithm works very well for both long and short answer grading due to it's incremental nature, which allows the system to be used for any kind of automatic grading system. The proposed system provides atleast 10% higher grading accuracy when compared with it's non-machine learning counterparts.***

Index Terms**: *Grading, language, processing, short, long, accuracy.***

I. INTRODUCTION

Automatic grading of student answers has been a field of research for professors and teaching assistants alike due to the advantages it possesses for both long term and short term usage. A well trained system can not only provide correct marking scheme, but also help in speeding up the process of result evaluation, which can save a lot of time and expenditure by the universities. Generally, the evaluation of answers is done by experts in the area, and experts are human beings, thus the grading is sometimes affected by human psychological factors such as the mood of the checker, the pressure on the checker in terms of time, the environmental conditions in which the grading is being done, and others. Even if the grading authority grades the papers accurately, the data entry done for these evaluations is managed by teaching assistants or lab assistants, who can make mistakes while adding the marks in the online systems, thus there is a lot of manual dependency in this field. In order to remove this dependency, there is a need of an accurate grading solution, which is described in this paper.

Usually, the process of grading uses natural language processing to extract action words from the actual answer, and the student answers. These action words are either nouns, pronouns, verbs, adverbs or adjectives. Once these action words are extracted, the synonyms are evaluated for each of the action words and are stored in the memory for further comparison. The actual answer words and it's synonyms are compared with each of student's answers and their synonyms, in order to evaluate a matching score. Based on this matching score the student marks are evaluated. In this process, a lot of mathematical computations are involved, for example, when matching the given answer words with the student answer words, there is a need of term frequency and inverse document frequency evaluation, whereas when converting the score into marks there is need of ratio evaluation, which involves certain computations. All these steps are explained in detail in the proposed work section of this text.

The next section demonstrates various algorithms for grading short and long answers, followed by the proposed approach for grading answers based on machine learning, and finally we conclude the paper with some finer observations and some future work which can be carried out by researchers in order to further improve the performance of the system.

II. LITERATURE SURVEY

Utilization of short answer preparing is displayed in [1], where a procedure reliant on irregular neural frameworks to get comfortable with the association between a paper and its doled out score, with no segment structuring is proposed. The results exhibit that this structure, which relies upon long transient memory frameworks, beats a strong standard by 5.6% to the extent quadratic weighted Kappa. For short and open answers (SOAs), the work in [2] presents a novel technique regarding learning assessment of SOA questions using thought maps. They propose a circumstance for using this kind of request in web learning conditions and depict their strategy for subsequently surveying this sort of request using thought maps and similarities measures. If per clients need to survey unmistakable systems for substance closeness, by then they can insinuate [3], where experts take a gander at and research two similarity measure techniques, cosine likeness and inactive semantic examination. The parameters that was used to evaluate the execution of the strategies are the computational unpredictability - assessed by the proportion of CPU and memory use, and page load time - and accuracy - evaluated by Pearson Correlation and Mean Absolute Error. The results exhibited that both estimation exhausted same proportion of memory. They furthermore ensure that cosine has a better server act so loved than be executed in e-learning customized article scoring structure.

Revised Manuscript Received on May 06, 2019

Simran Agrawal, persuing Master of Technology degree in Computer Science and Engineering from Shri Ramdeobaba college of Engineering and Management, Nagpur, India.

Dr. Avinash J. Agrawal, Associate Professor in Shri Ramdeobaba college of Engineering and Management, Nagpur, India.

For long answers, experts in [4], developed a Two-Stage Learning Framework (TSLF) which organizes the upsides of both part planned and all the way AES methods. In tests, they examine TSLF against different strong baselines, and the results display the ampleness and generosity of the models. TSLF outflanks all of the baselines on five-eighths of prompts and achieves new top tier ordinary execution when without negative precedents. In the wake of adding a few papers to the first datasets, TSLF pulsates the features planned and from beginning to end baselines, everything considered, and exhibits exceptional generosity. Article evaluation takes a long time at whatever point changed physically. Therefore, investigates on modified article scoring have been growing rapidly starting late. The methodology that is commonly used for customized work scoring is Cosine Similarity by utilizing pack of words as the part extraction. In any case, the component extraction by using pack of words did not consider to the solicitation of words in a sentence. Meanwhile, the solicitation of words in an article has a fundamental occupation in the evaluation.

In [5], a modified paper scoring system subject to n-gram and cosine comparability was proposed. N-gram was used for feature extraction and changed to part by word instead of by letter with the objective that the word solicitation would be considered. In light of evaluation results, this structure got the best association of 0.66 by using unigram on request that don't consider the solicitation of words in the fitting reaction. For request that consider the solicitation of the words in the fitting reaction, bigram has the best relationship regard by 0.67. In [6], experts have developed a module where understudy response and right answer will be set up by at first isolating them into token for instance words. Later on thing articulation and activity word social affair will be alloted to each and every word with the help of Part-Of-Speech (POS) tagger. This endeavor is developed by NLP framework. Each and every outflow of understudy response is differentiated and right answer. In case clear match is found in word similarly as POS tag and word position in sentence the scores are doled out. After score task Final scores are dictated by making summation of selected scores everything being equivalent. This procedure is valuable for both long and short answers.

In [7], piece based model is proposed with the name "Goodness" which uses a conventionality score to find the advantage of organizing and subject to this score the exploring is done. Multifaceted nature measures, like the Flesch Reading Ease Score (FRES) are furthermore used as features in the system. Despite those, parse tree features, like the typical parse tree significance and the amount of subordinate stipulations (SBAR) in the substance are used for score evaluation. This methodology is helpful for appraisal of long messages.

In [8], researchers present an other unsupervised technique which deals with understudies' answers extensively using substance to content equivalence. Unmistakable String-based and Corpus-based resemblance measures were attempted freely and a short time later joined to achieve a biggest association estimation of 0.504. The practiced relationship is the best regard achieved for unsupervised procedure Bag of Words (BOW) when stood out from some standard systems. While, [9] is revolved around the improvement of modified short answer scoring. Some customized scoring systems used on long answer have shown perfect results in giving a score on the understudies answer. Customized long answer structures use the information recuperation method to measure similarity between understudies answer and references answer. Modified short answer scoring does not give the best result yet. Short answer has a limited word in each answer. Each answer contains one articulation to three sentences. Examination of the short depiction that has set number of words requires exceptional dealing with, especially in the weighting methodology. With the controls of the path toward weighting the word, it is inconceivable with repeat appear, in light of the fact that the words occasion is exceptional. This examination endeavors to consider a couple of methods that apply the covering procedures to choose the dimension of closeness between the references answer and understudies answer. From the examination it shows that the procedure Cosine Coefficient has ideal results over the Dice and Jaccard Coefficient systems. In [10], the investigators explored the data used in Automatic Short Answer Grading (ASAG) investigate, contemplating various points of view from the language to number of request, answers, etc. By then, they looked ordinary language taking care of and AI methodologies are the most used in the field. Starting now and into the foreseeable future, they showed the focal point of the examination, how answers are shown in order to remove features that can envision their scores. In conclusion, they showed how investigators surveyed their structures and how they can (or can't) be stood out from each other. All presented results exhibits the substance and advancement of ASAG research using AI procedures. Open datasets are available for an actually lengthy timespan earlier, and inspect in the field is accessible to new strategies, datasets and uncommonly to significant understanding, that has been starting late mixing it up of areas and still very underexplored in ASAG. In [11], a philosophy on long answer evaluation using lexical and semantic similarity measure has been presented. The goal of this work is to display a structure which naturally surveys the long answers from the examinee and hereafter decrease the time and effort of human intervention similarly as make the appraisal technique impartial to the entire customer. In this work, first the customer answers (examinee answer) are facilitated with standard answers (investigator answer) using lexical likeness measure. In testing stage, five courses of action of request answers have been seen as where each set contains a lone request from a subject space and its five extraordinary answers. The system ardent a commendable exactness as shown by human decision. In the accompanying time of the work, both the proper reactions have been dissected using semantic equivalence measure. In this stage, the synonymous articulations of the watchwords from both the fitting reactions are recouped from the semantic dictionary WordNet to grow the needful and critical spread between the suitable reactions. Applying this semantic resemblance evaluating system on a comparable request answer sets, the exactness of appraisal has been extended which is affirmed by a master.

In [12], the researchers proposed an approach to manage check the dimension of learning of the understudy/understudy, by surveying their entrancing test answer sheets. By addressing the connecting with answer as outline and differentiating it and standard answer are the key walks in our strategy. The understudy's unmistakable answer and standard answer is changed over into its graphical structure and a while later, to apply a bit of the likeness gauges, for instance, string match, wordNet and spreading process for the calculation of equivalence score are the genuine walks in the proposed count. The count gives a response for the automation of realistic answer appraisal process.

III. PROPOSED WORK

The proposed approach can be demonstrated with the help of the following steps,

- Natural language processing to extract action words and perform stemming
- Evaluating the synsets for the standard answer
- Division of student answers into training and testing sets
- Learning from the training set via Q learning mechanism
- Application of learned weights on the testing set

We prepared a dataset which had the question, the actual answer, student answers and the marks given by the examiner manually for each of the student answers. These marks served as a training set for the algorithm. In the first step, the given answer and the student answers are processed through a natural language processing application programming interface (API), which evaluates the answers and stores their action words in separate lists. Let the list of words for the given answer be Lg, and the list of words for the students answers be Ls1, Ls2, Ls3 ... LsN, where N is the number of student answers in the dataset.

Each of the words in the Lg list is then given to a synset evaluation unit, where the synonyms are evaluated and stored in another list, named LgS. This list is further used for processing and scoring purposes. The following machine learning algorithm is then applied on the lists,

- Initialize the machine learning parameters, namely,
  - Min threshold = MTh, which defines the minimum number of correct words to be identified in the student answer, so that the student can get maximum marks
  - Machine learning Training Threshold = Tth, which is the number of answers to be used in the training set
  - Error diff val, Ed, which defines the threshold difference of the marks given to the student by the algorithm to the actual marks given to the student
- Initialize the machine learning factor (Fml) as 0
- For each ith answer in the training set, evaluate the term frequency from between each word in the answer from the list Lsi, and the list LgS, let this term frequency be called Tfi
- Sum up the Tfi values, and divide them by the number of words in the list Ls, to find the answer score, using the following equation.
  Score (Si) = Sum(Tfi) / Words in Ls
- For all the training answers, sum up this score Si, and evaluate it's mean, let this mean value be Msi
- For all the testing answers, evaluate the score using the term frequency method, and then multiply the score with Msi, to get the final score for each of the testing set answers

The process is repeated for all the training and testing set answers, due to which the Msi value auto tunes itself and obtains the final true score which has to assigned to the user. The result and analysis of the technique is given in the next section.

IV. RESULT AND ANALYSIS

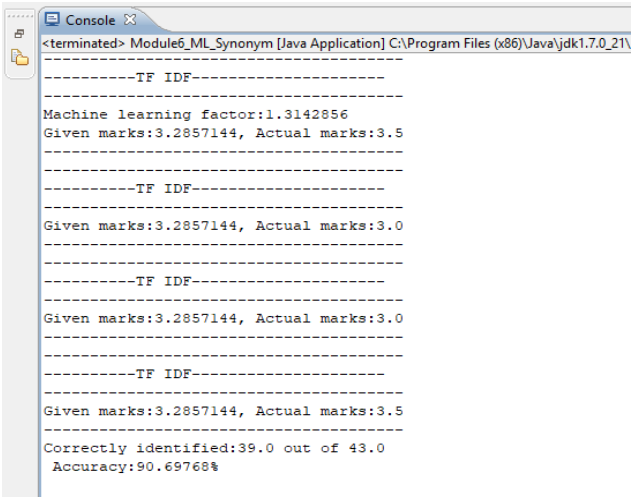
We analyzed the system on both long and short answer datasets. The long answer datasets were prepared manually from the previous year's evaluation papers accumulated at our institute, while for short answers we took the online datasets available at kaggle. The following results were obtained in terms of accuracy of grading,

Table 1. Obtained Result for both short answer and long answer dataset

			Module 2	Module 3	Module 4	Module 5	Module 6
Short Answer Dataset	With Stemming	Assign 1	18.96%	32.75%	35.7%	39.65%	67.85%
		Assign 2	11.42%	22.85%	18.75%	23.33%	31.2%
		Assign 3	5.52%	26.26%	17.46%	20.73%	50.79%
	Without Stemming	Assign 1	4.31%	37.93%	50%	38.79%	82.14%
		Assign 2	8.57%	31.90%	18.75%	40.95%	31.25%
		Assign 3	11.05%	35.94%	30.15%	45.16%	41.26%
Long Answer Dataset	With Stemming		32.55%	68.21%	46.51%	44.96%	76.74%
	Without Stemming		27.13%	68.99%	74.41%	41.86%	90.69%

From the table 1, the module 2 is where we applied normal scoring using Tf, while module 3 used both Tf and Idf terms, module 4 is the output of pure machine learning, module 5 is module 2 but with synonyms included, and finally module 6 is the proposed system output. From the result analysis we can evaluate that the proposed system has very high accuracy for long answer sets, while it has good accuracy for short answer sets. Short answer sets usually do not have enough comparison data, and students might write answers which do not match in entirety with the input standard answers, thus the accuracy for short answers is not as high as that for the long answer counterparts.

We evaluated the results using a Java based developed code, and the following screenshot determines the output for it,



From the above screen we can see that the Machine learning factor has been optimized, and the accuracy for the set is evaluated.

V. CONCLUSION

From the results, we can observe that the accuracy of prediction for the proposed machine learning based system is higher by almost 20% when compared with traditional language processing techniques, for both long answers and short answers alike. We also observe that the proposed system can be used for any domain answers, and performs very well provided the given training set has been tuned properly.

FUTURE WORK

As machine learning and AI are gathering a lot of momentum, so the researchers can further enhance the accuracy of the system by using modern AI techniques like deep nets and evaluate it's performance on various long and short answer sets in order to further enhance the system accuracy.

REFERENCES

1. Mahana, Manvi, Mishel Johns, and Ashwin Apte. "Robotized exposition reviewing utilizing AI." Mach. Learn. Session, Stanford University (2012).
2. Magooda, Ahmed Ezzat, et al. "Vector based systems for short answer evaluating." The Twenty-Ninth International Flairs Conference. 2016.
3. Liu, Jiawei, Yang Xu, and Lingzhe Zhao. "Robotized Essay Scoring dependent on Two-Stage Learning." arXiv preprint arXiv:1901.07744 (2019).
4. Patil, Shweta, and SonalPatil. "Canny mentoring framework for assessing understudy execution in illustrative answers utilizing characteristic language preparing." International Journal of Science and Research 3.9 (2014): 779-783.
5. Pribadi, FeddySetio, et al. "Programmed short answer scoring utilizing words covering techniques." AIP Conference Proceedings. Vol. 1818. No. 1. AIP Publishing, 2017.
6. Kaur, Amarjeet, et al. "Calculation for Automatic assessment of single sentence spellbinding answer." International Journal of Inventive Engineering and Science 1.9 (2013): 112-121.
7. Galhardi, Lucas Busatta, and Jacques DuílioBrancher. "AI Approach for Automatic Short Answer Grading: A Systematic Review." Ibero-American Conference on Artificial Intelligence. Springer, Cham, 2018.
8. Taghipour, Kaveh, and HweeTou Ng. "A neural way to deal with robotized paper scoring." Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016.
9. Kaur, Amarjeet, and M. Sasikumar. "A near investigation of different methodologies for mechanized evaluation of unmistakable answers." 2017 International Conference on Computational Intelligence in Data Science (ICCIDS). IEEE, 2017.
10. Nandini, V., and P. Uma Maheswari. "Programmed appraisal of clear answers in online examination framework utilizing semantic social highlights." The Journal of Supercomputing (2018): 1-19.
11. Contreras, Jennifer O., ShadiHilles, and Zainab BintiAbubakar. "Robotized Essay Scoring with Ontology dependent on Text Mining and NLTK apparatuses." 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE). IEEE, 2018.
12. Fauzi, M. Ali, et al. "Programmed Essay Scoring System Using N-Gram and Cosine Similarity for Gamification Based E-Learning." Proceedings of the International Conference on Advances in Image Processing. ACM, 2017.

AUTHORS PROFILE

Simran Agrawal received Bachelor of Engineering Degree in Computer Science and Engineering from Shri Balaji Institute of Science and Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya University, Betul, India 2016 and persuing Master of Technology degree in Computer Science and Engineering from Shri Ramdeobaba college of Engineering and Management, Nagpur, India.

Dr. Avinash J. Agrawal received Bachelor of Engineering Degree in Computer Technology from Nagpur University, India and Master of Technology degree in Computer Technology from National Institute of Technology, Raipur, India in 1998 and 2005 respectively. He completed his Ph.D. from Visvesvaraya National Institute of Technology, Nagpur. His research area is Natural Language Processing and Databases. He is having 20 years of teaching experience. Presently he is Associate Professor in Shri Ramdeobaba college of Engineering and Management, Nagpur,India.