TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 3595-3600

Intelligent Scoring Systems for Descriptive Answers-A Review

[1]P. Sree Lakshmi, [2]Kavitha
[1]Research Scholar, [2]Associate Professor, School of CSA, REVA University, Bengaluru

Article Info
Volume 83
Page Number: 3595-3600
Publication Issue:
May-June 2020

Article History
Article Received: *19 August 2019*
Revised: *27 November 2019*
Accepted: *29 January 2020*
Publication: *12 May 2020*

Abstract

Evaluation is one of the vital components in the teaching-learning process as it helps in periodically assess and modify the teaching activities according to the learner requirements. In case of descriptive answers, it is the most time consuming and error prone process. The proliferation of techniques in Deep learning and Natural language processing helps to automate the task of Evaluation, so that the evaluator can spend crucial time in improving the teaching learning process. In this paper, a review over various automatic short answer grading methods which includes the methods used for feature selection, the datasets used ,various text similarity methods applied, issues addressed has been explored. This survey helps in getting awareness on different approaches used for designing ASAG systems and also its drawbacks and enhancements required has been discovered.

Keywords: *Automatic grading, Short descriptive answer, Natural language processing.*

1. Introduction

One of the famous quotes of Nelson Mandela says that, **"Education** is the most powerful weapon which you can use to change the world". India is the fastest developing country in which the quality of education plays a vital role for its promising future. In Teaching-learning process, evaluation is utmost important phase because it helps in diagnosing the weakness of the students, further helps in identifying the areas where improvement is required and ultimately helps to realize the goals more efficiently. The Questions used for evaluation can be mainly categorized into two types: i) Objective questions, where the students will select the correct response from several alternatives or supplies a word or short phrase to answer a question or complete a statement. Examples of Objectives type questions are Multiple choice questions, true or false (yes or No), fill in the blanks, Matching and ii) Subjective questions, or essay, where the student has to organize and present an original answer. Examples for Subjective type questions are short-answer essay, extended-response essay, and problem solving etc.

Evaluation of Objective type questions is easy when compared to Subjective Questions. In order to test the knowledge of the student both objective and subjective questions should be included in the question paper because subjective questions helps to evaluate the students understanding of subject and its concepts, to Understand the thinking and problem-solving ability of the candidate and Evaluates the writing and communicating ability of the candidate.

The manual evaluation of descriptive answers is very tedious task. It requires lot of time and human power. Because of mood swings of the evaluator it is an error prone process and different evaluators will assign different marks for the same answer. Many issues occurred like nearly 15 students committed suicide in Telangana intermediate board in April 2018 and every year issues like suspension of evaluators happening and moreover students are spending lot of money for revaluation.

We can overcome above problems by automation of evaluation process. Most of research happened on automatic evaluation of Multiple Choice Questions (MCQs) and questions with true/false because answers to such questions can be automatically evaluated by using Computer Assisted Assessment (CAA) systems. Still more research is needed to solve the issues regarding automatic evaluation of handwritten descriptive answers.

The Research on automatic assessment of answers till date can be mainly categorized into two types:1) Automatic *Essay grading (AEG)*, where more focus will be on Writing style, grammar and consistency of the text(Higgins et al. 2004; Pérez 2004) [15][16]and 2)

May-June 2020
ISSN: 0193-4120 Page No. 3595-3600

Automatic *Short Answer grading* (ASAG) where focus is on semantic content of the answer rather than general style (Burrows et al. 2015)[3].

We are mainly focusing on automatic grading of short answers which will have features like in this type of questions the answer cannot be recognized from the question but requires thinking of external knowledge, the response is given in natural language, the answer length should be between one phrase or a paragraph and while evaluation the focus will be on content than writing style.

Automatic short answer grading (ASAG) is the process of assessing the descriptive answers automatically using Natural Language processing techniques. It can be described as either a *Regression task*ie assigning some grade/score for the respective answer or a *Classification task*i.e. assigning some label like correct, partially correct, incorrect etc. to the answer.

In this paper a comparative analysis on various automatic short answer grading methods is done. The paper is organized as follows: Section 2 presents an overview of related work in the field of ASAG. Section 3 explains proposed method and Research challenges identified and finally section 4 presents the conclusion.

2. Related Works

According to Burrows et al. (2015) [3] the earlier research in ASAG can be mainly categorized into 4 approaches.

1)Concept mapping methods: In this method, the student answer is considered to be made up of several concepts and while grading the presence or absence of the concept is detected and concept level mapping is stated at sentence level. In these phase deep and complex NLP techniques are used to extract syntactic and sematic representations of student answers. Example questions are like questions that asks both solutions and clarification to a problem, questions that queries several reasons for the same problem.

Manning and Schütze(1999)[1] had discussed constituents and syntactic dependencies among texts using syntactic analyzers.

Burstein et al.(2001)[6] explored on analyzing the discourse structures of texts through rhetorical parsers.

C-rater(Concept Rater)Leacock and Chodorow (2003)[13] considered that student answers consist of multiple concepts and they are accordingly matched the concepts of student answers with those of teacher's answers using techniques like anaphora, syntactic variation, morphological variation, spelling correction and so on.

2) *Information Extraction methods*: These methods use a series of pattern matching methods like regular expressions, parsing trees for evaluation. These methods help to extract structured data from unstructured resources. For example AutoMark (Mitchell et al. 2002)[8] where both student and teacher answers are represented in the form of Parse tree, and applied pattern matching as an information extraction task. Auto-marking (Sukkarieh et.al 2003) [14] designed the training set using two hand crafted patterns for each question. The observed that this approach is more effective than a k-nearest neighbor baseline with bag-of-words features weighted by TF-IDF.

3) Corpus based methods. They exploit the statistical properties of large document corpora while grading which helps in interpreting the synonyms in short answers.

4) Machine learning methods these methods use the measurements extracted from Natural language processing techniques and later they are combined into single grade or score using classification and regression model. Features may involve bag-of-words, n-grams etc. and learning algorithms includes decision trees, SVM (support vector Machines), Naïve Bayes and K nearest neighbors etc.

Neethu George et.al [2] Proposed D-DAS Model (Deep Descriptive Answer Scoring model) for automatic essay evaluation using deep learning and natural language processing. It consists of embedding layer, LSTM-RNN layer, dropout layer and dense layer. Here the entire answer is given as input to the embedding layer, which converts it into a glove vector representation. Then LSTM-RNN will learn temporal data from embedding layer and the final glove vector which is the semantic representation of the entire answer is given as input to the dropout layer, where in regularization technique for handling overfitting of data is applied. Finally, Softmax activation function is used in dense layer that assigns one hot encoded score for each answer.

Piyush Patilet.al[4] Proposed a model that uses machine learning and NLP to evaluate subjective answer. Tasks like Tokenizing of words and sentences, Part of Speech tagging, Chunking, Chinking, Lemmatizing words and Word netting are performed and semantic meaning of the context is also considered while evaluation. A student written answer from an answer sheet is scanned ,performs preprocessing and extracts text from the answer. These are compared with model answers which contains keywords, Grammar and QST(Question Specific Things).

MdGulzar Hussainet.al[5] Has proposed the model for automatic evaluation of descriptive answers written in Bangla language. Various steps includes Keyword generation, Based on Keywords the answers are searched and extracted from both closed domain (Evaluator provided answers) and Open domain(Wikipedia, web pages ,Blogs, World Wide Web etc.).Proposed algorithms for comparison of answers and for analyzing grammatical and spelling mistakes.

Sonakshi Vij[7]is the first model that applies WordNet Graphs for text similarity in Automatic short

May-June 2020
ISSN: 0193-4120 Page No. 3595-3600

answer assessment. Evaluation is done by identifying the common nodes between node set of Ideal answer and Student answer WordNet graphs. Evaluation is done by considering the semantic and structural dependencies notions of text similarity.

Wallace Dalmetet.al[9] Had applied Siamese Manhattan LSTM, which is a deep neural network to find the semantic similarity between two answers. Google's word2vec was used to convert answers into matrix form. Sentences are first converted into fixed length vectors which are fed into embedding layer where all word embedding and they are represented in a matrix form. These Two embedded matrices which contains answers to be compared are given as input to LSTM network to capture semantic similarity and finally similarity function scores between 0 and 1.

3. Proposed Method and Research Issues

After analysis of the literature on ASAG systems, the first thing to be done for designing a model for automatic grading is first understand the text, identifying the best features using Feature extraction methods, with the help of these extracted best features comparison of model and student answers is done by using text similarity approaches, then based on similarity scores are assigned. Then finally, human and system generated scores are compared to test the accuracy of the model.

Some of the research issues identified are as follows:

1) As different interpretations of Rubrics are used there will be inconsistency in the systems.
2) For better accuracy more training models are required.
3) Most of the automated systems developed are not providing feedback to student, which is more important for knowing the understanding level for student.
4) The automated systems are more prone to cheating.
5) Most of the systems are domain specific. Systems that are able to grade any subject should be developed.
6) Systems should be able to grade the answers where a part of answer is in one place and remaining in other i.e. which are written in different places also.
7) Scalable automatic grading systems need to be developed which can be used based on requirement.
8) Most of the automatic grading systems developed are not for handwritten answer evaluation. Still more research is required in this area.
9) Systems should be designed in such a way that it should grade answers having figures and equations, scratched lines, improper word, character spacing and when text is highlighted using boxes.

Table 1: Analysis of various Automatic Grading Systems in literature

Sl. no	Author	Proposed model	Method/ Techniques	Datasets	Measures	Findings	Remarks/Conclusion
1	Neethu George et.al[2]	D-DAS Model, (Deep Descriptive Answer Scoring model	Natural language processing, Deep learning, LSTM-RNN Regularization Technique- dropout ,Soft max activation function	Dataset is prepared manually from 50 distinct answer scripts. 50 iterations and 100 epochs were used for training and experimented on 500 epochs	Kappa score	For 90% training data, the accuracy obtained by the D-DAS model using simple LSTM is 83%, using deep LSTM is 82% and bidirectional LSTM is 89%.	D-DAS Model with Bi-LSTM proved more performance than other neural network based models
2	PiyushPatil et.al [4]	Subjective answer evaluative by considering length of answer, QST, keywords.	Machine Learning, NLP Tokenizing words and sentences, Part of Speech tagging, Chunking, chinking, Lemmatizing words and Word netting, Naïve Bayes	3 questions for 20 students, 21 are used for training dataset	Cosine similarity, Fuzzy-wuzzy in fuzzy Logic	21 are used or for training	ML techniques gave satisfactory results. The accuracy of the classifier can be improved by feeding huge amount of training data.

May-June 2020
ISSN: 0193-4120 Page No. 3595-3600

			classifier				
3	MdGulzar et.al[5]	proposed a basic method to evaluate Bangla subjective answer	Corpus based method Keyword matching, linguistic analysis	20 questions for 10 students	Absolute error, Relative error	The Relative Error is 10% for the proposed model.	Mechanism for checking Synonyms of the words need to be implemented. Proposed model should be implemented using machine learning algorithms for better accuracy. Inputs from handwritten answers should be considered.
4	Sonakshi Vij [7]	Machine Learning model	Machine learning ,WordNet Graphs for Text Similarity	400 students answer sheets of Social studies	RMSE, Root Mean Square Error for comparison	RMSE=0.39	Proposed model will not suite well for Technical subjects like computer science and Engineering as the Word Net don't contain all technical words and definitions. The WordNet graph for Misspelled words will not be generated.
5.	Wallace Dalmet et.al[9]	Applied Manhattan LSTM for Text similarity	Deep Learning, Siamese Manhattan LSTM deep neural network for Text Similarity Google's word2vec	Two embedding matrices which consists of the answers to be compared.	Similarity function	Siamese Manhattan LSTM outperforms other methods of sentence similarity such as cosine similarity and word mover's distance	Deep learning methods can be applied to make the system more efficient. While Training, by swapping the words with their synonyms, Data augmentation can be applied to generate new pairs of question-answer The system can be mounted according to the requirement of examiner by scaling the output of similarity function
6	Yuan Zhanget. al[10]	Student and domain/question models are used for improving the performance of ASAG systems.	Student and domain/question models, Deep learning model, DBN(Deep Belief Network), stop word removal , tokenization ,punctuation removal and word correction	Cordillera, a natural language tutoring system. 150 students Participated.	Accuracy, Area under the curve (AUC), Precision, Recall, F-measure	DBN outperforms all five classic ML classifiers on accuracy, AUC, precision and F-measure. Only on recall, DBN performs slightly worse than SVM.	Instead of considering only answer model at the time of evaluation, the inclusion of student and question models enhanced the grading performance. Proposed model is tested only on one dataset. Evaluations on largescale datasets are required.KC's are drafted manually in a Qmatrix , but better if it would be done automatically. This framework predicts answers only in two ways as correct or incorrect. We can design a framework that predicts various categories of incorrect answers.

May-June 2020
ISSN: 0193-4120 Page No. 3595-3600

7.	Vijay Rowtula et.al[11]	Self-supervised Model that evaluates handwritten descriptive answers by including semantics.	CNN,Information Retrieval and Extraction (IRE) and, Natural Language Processing (NLP),feature based word spotting	Class Room Dataset (CRD): 6 students answered by 96 students Controlled Dataset (CD), SciEntsBank Dataset (SE):	precision, recall and F1-score	This framework integrates ideas from IRE and NLP and word spotting. On CRD dataset it has correlated more with human evaluator.	This model does not suites for answer scripts that have figures and equations, scratched lines, improper word, character spacing, when text is highlighted using boxes.
8	Tianqi Wang et.al[12]	Considering Rubrics for ASG	Neural networks,Glove embedding	ASAP-SAS dataset, 1,704 training data, 522 test data	MSE, RMSprop optimizer	Inclusion of Rubric component in state of art neural SAG models improves the performance.	Instead of word level attention, by considering context the association between answers and key elements can be improved. Other categories of rubrics can also be explored.

4. Conclusion

Evaluation of descriptive answers is still a manual task, which is more time consuming and error process. To solve this issue, Many automatic short answer grading systems have been developed but still the accuracy and precision need to be improved compared to human evaluator. Initially deep NLP techniques like syntactic analyzers, Rhetorical parsers and semantic analyzers were used but as a short student answer will not provide sufficient lexical features for analysis, they were not applied to the maximum extent. Later shallow NLP techniques along with machine learning were used. Many approaches are used like concept mapping, information extraction, Corpus based methods and finally Machine learning models like LSTM, WordNet graphs, CNN, RNN, DBN etc. were used. Mainly used measures for accuracy are MSE, Precision, Recall, F1 score, RMSE etc.

References

[1] Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing, vol. 999. MIT Press, Cambridge (1999).

[2] Neethu George, Sijimol PJ, Surekha Mariam Varghese ,Grading Descriptive Answer Scripts Using Deep Learning International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-5 March, 2019.

[3] Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. Int. J. Artif. Intell. Educ. 25(1), 60–117 (2015).

[4] Subjective Answer Evaluation Using Machine Learning, PiyushPatil ,SachinPatil, Vaibhav Miniyar ,Amol Bandal ,International Journal of Pure and Applied Mathematics, Volume 118 No. 24 2018 ISSN: 1314-3395.

[5] Assessment of Bangla Descriptive Answer Script Digitally, MdGulzar Hussain∗, SumaiyaKabir†, Tamim Al Mahmud‡, Ayesha Khatun§, MdJahidulIslam, International Conference on Bangla Speech and Language Processing(ICBSLP), 27-28 September, 2019.

[6] Burstein, J., Leacock, C., Swartz, R.: Automated evaluation of essays and short answers (2001).

[7] A Machine Learning Approach for Automated Evaluation of Short Answers Using Text Similarity Based on WordNet Graphs, Sonakshi Vij · Devendra Tayal · Amita Jain, Wireless Personal Communications ,© Springer Science + Business Media, LLC, part of Springer Nature 2019 https://doi.org/10.1007/s11277-019-06913-x.

[8] Mitchell, T., Russell, T., Broomhead, P., Aldridge, N.: Towards robust computerised marking of free-text responses (2002).

[9] Wallace Dalmet, Abhishek Das, VivekDhuri, MoinuddinKhajaand Sunil H. Karamchandani, Siamese Manhattan LSTM Implementation for Predicting TextSimilarity and Grading of Student TestPapers,© Springer Nature Singapore Pte Ltd. 2020H. Vasudevan et al. (eds.),Proceedings of International Conference on Wireless Communication, Lecture Notes on Data Engineering and Communications Technologies 36,https://doi.org/10.1007/978-981-15-1002-1_60.

[10] Going deeper: Automatic short – answer grading by combining student and question models

May-June 2020
ISSN: 0193-4120 Page No. 3595-3600

,Yuan Zhang ,Chen Li1, Min Chi, User Modeling and User-Adapted Interaction https://doi.org/10.1007/s11257-019-09251-6,© Springer Nature B.V. 2020.

[11] Vijay Rowtula ,Subbareddyoota, C. V . Jawahar , Towards Automated Evaluation of Handwritten Assessments, 2379-2140/19/$31.00 ©2019 IEEE DOI 10.1109/ICDAR.2019.00075, 2019 International Conference on Document Analysis and Recognition (ICDAR).

[12] Tianqi Wang, Naoya Inoue, Hiroki Ouchi, Tomoya Mizumoto, Kentaro Inui, Inject Rubrics into Short Answer Grading System, Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo), pages 175–182 Hong Kong, China, November 3,2019 Association for Computational Linguistics https://doi.org/10.18653/v1/P17.

[13] Leacock, C., Chodorow, M.: C-rater: automated scoring of short-answer questions. Comput. Humanit. 37(4), 389–405 (2003).

[14] Sukkarieh, J.Z., Pulman, S.G., Raikes, N. (2003). Auto-marking: using computational linguistics to score short, free text responses. In Proceedings of the 29th annual conference of the international association for educational assessment (pp. 1–15). Manchester.

[15] HigginsD., Burstein,J., Marcu,D., Gentile,C.: Evaluating multiple aspects of coherence in students says. In: HLT-NAACL (2004).

[16] .Pérez,D. : Automatic evaluation of user's short essays by using statistical and shallow natural language processing techniques. Advanced Studies Diploma (Escuela Politécnica Superior ,Universidad Autónoma de Madrid) (2004).