IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Vector Similarity Measure for ASAG

Chandralika Chakraborty[1]*, Udit Kr. Chakraborty[2], Bhairab Sarma[3]

[1]Dept. of Information Technology, Sikkim Manipal Institute of Technology, Sikkim Manipal University, India
[2]Dept. of Computer Science & Engineering, Sikkim Manipal Institute of Technology, Sikkim Manipal University, India
[3]Dept. of Computer Science, University of Science & Technology Meghalaya, India

Corresponding Author: chandralika.c@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i4.959963 | Available online at: www.ijcseonline.org

Accepted: 20/Apr/2019, Published: 30/Apr/2019

Abstract: Automated Short Answer Grading (ASAG) has been an area of active research for quite some time now. Several theories and implementation have come up, but a stable method, suitable for all genres of answers is yet to be standardized. The most accurate results for short answer grading have been found for substantially longer texts which have scope for information retrieval. Smaller answers however suffer on this front and have been a bottleneck of sorts. This paper presents a simple method to evaluate very short answers, using cosine similarity method between students' answers and model answers prepared by subject experts. The proposed method is simple, fast and easy to implement and returns scores having fair correlation with human evaluated scores.

Keywords: *ASAG, Students Answer, Model Answer, Cosine Similarity.*

I. INTRODUCTION

Evaluating answer scripts of students is a demanding task. A single teacher evaluating a large number of students, may not always reflect the correct judgement. In countries like India, with teacher –student ratio being as high as 1:20 in technical courses [1], even the time allotted for evaluation is insufficient[2].

The problem is further augmented by the fact that evaluators differ in their approaches to evaluation and may adhere to self defined standards. While some may be overtly strict, other may be lenient, resulting in differences in scoring. While techniques of standardization of evaluation using answer schemes or rubrics do exist, there preparation needs training and time [3]. Moreover, scoring rubrics, however detailed they may be are not always sufficient to safeguard the interest of the learner in the face of human inconsistency. The process is also time consuming and eats up resources which could otherwise have better uses.

A solution to these problems lie in automating the process of answer evaluation. Computer based Assessment Systems (CbAS) has been an active area of development for sometime now. However, the types of questions popularly used in CbAS used in e-learning are multiple choice types (MCQ's) or one with short answers [4][5]. The popularity of MCQ's may be attributed to some of its features like objectivity, user-friendliness, quantifiability, and as well as the fact that they provide scope for more effective and efficient feedback [6]. However, there are serious limitations with MCQ's as these fall short of mark when the learners' theoretical knowledge has to be tested[9][10]. Furthermore, MCQ's have the serious problem of being unable to check guess work. An MCQ with four options presents a one in four chance of 'guessing' the correct answer. This can effectively result in a person having 20% correct answer knowledge of a subject scoring 40% marks in a paper having MCQ's with four options [7].

Such issues of e-learning evaluation and MCQ's may be handled through the incorporation of text based evaluation. The current paper presents a sentence similarity based method for evaluation of short text based answers, otherwise also known as Automatic Short Answer Grading (ASAG). The answers handled in this paper are restricted in length to a maximum of three sentences, the average length of the answers being 70 words. While such answers may be classified as short answers, an ideal length for a short answer may be around 150-250 words, as considered by most universities [8][9]. A short answer, viewed from the perspective of evaluation has to meet the five point criteria [10]. These are:

1. the question must require a response that recalls external knowledge instead of requiring the answer to be recognized from within the question.
2. the question must require a response given in natural language.
3. the answer length must be restricted.
4. the assessment of the responses should focus on the

content instead of writing style.
5. the level of openness in open-ended versus close-ended responses should be restricted with an objective question design.

While the list does not explicitly specify the length of an answer or even a boundary to qualify as a short answer, the assessment needs to be focussed on content. The average length of the answers under consideration being 70 words, the short length makes the evaluation of such answers substantially difficult using computer based techniques as algorithm based discovery and extraction of knowledge requires a sizeable data [11].

The current work presents a method of evaluating what may be considered as very short answers, which are classified as Text-Explicit (TE) and do not require inference to be drawn between two or more text segments [12]. Pedagogically, such answers are significant as the learner is not allowed the flexibility to answer elaborately and the answers evaluate knowledge summarization. The learner has to learn the defining features of an idea, present the essential concepts and their interrelationships and can be quick and holistically evaluated [13].

The paper is organized as follows. Section I contains the introduction to Automatic Short Answer Grading and Section II presents a brief review of work done for automated grading. Section III explains the proposed methodology for evaluating the text answers, while, Section IV reports the results of detailed experiments carried out. Section V concludes the paper with future directions for research, which is followed by references.

II. RELATED WORK

A comprehensive survey on the topic by Burrows et al [10] brings out that, "Research in grading natural language responses with computational methods has a history dating back to the early work of Page (1966)". Over the decades, various approaches have added to the repository of evaluative techniques used on short answers for automated grading. These include, but are not limited to concept mapping, information extraction, corpus based or machine learning methods.

The concept or facet mapping, initially proposed by Burstein et al[14] and later refined and/or reused by a number of researchers, led to some popular systems being developed. Information extraction, similarly led to its share of popular products being developed. These largely depended on regular expression or parse tree based pattern extraction[15].

The corpus based or machine learning methods, used differently by researchers like Lin [16] and Mihalcea [17] among others used corpus based features and relied on the word distributions and its variants like n-grams or bag-of-words.

The performance of all approaches vary based on the kind of answer, and one solution suitable for all forms of text is yet to be found. While the efficacy of the proposed systems continue to improve for longer versions of text based answers, the shorter ones pose challenges, reason being difficulty in information, concept or facet extraction from a smaller corpus size.

The work presented in this paper tries to bridge this gap and develop an evaluation technique for text based answers of very short length.

III. METHODOLOGY

The presented work, part of a bigger task, which aims at developing a technique to automatically evaluate text based answers, considers very short answers. The answers written by students are evaluated with respect to model answers created by subject experts.

Data Set:
The restriction being very short answers, the required dataset was created using answers by undergraduate students of Engineering. The total set consisted of three questions, posed to a group of seventeen (17) students. The answers had an average length of seventy (70) words and was evaluated by two evaluators separately, having been given three model answers to base their evalaution. The same answers were again system evaluated with respect to the same model answers. Table 1, lists the questions along with the model answers, while Table 2 shows some sample answers. The complete dataset has been made available online [18] for reference.

Table 1: List of Questions and Model Answers

Type	**Description**
Question 1	What is a program?
Model Answer 1	A program is a sequence of instructions to solve a particular task, using a programming language.
Question 2	What is an algorithm?
Model Answer 2	An algorithm is a finite set of steps carried out to perform a particular task in a finite amount of time.
Question 3	What is a flowchart?
Model Answer 3	A flowchart is a diagramatic representation of the steps of an algorithm, using specific symbols.

Though ideally expected to be a single sentence long only, some answers are longer. No preprocessing was done on any answers and they were treated as submitted by the students.

Table 2: List of Student Answers

Question No.	Student No.	Answer
1	1	A computer is a sequence of instructions for performing a task designed to solve a specific problem. Each program instruction is designed to be executable by a computer.
1	2	Program is a set of controlled instruction which solves a problem.
2	1	An algorithm is a step by step method of solving a problem. It is commonly used for data processing calculation and other related computer and mathematical operations.
2	2	Algorithm is a sequence of instructions to an unambiguous problem in a finite steps.
3	1	A type of diagram that represents an algorithm, workflow or process.
3	2	Flowchart is a visual representation to a sequence of instructions to an unambiguous problem.

Text Cosine Similarity:
The present implementation uses cosine similarity as the basis of evaluating the students textual responses.

Cosine similarity is a similarity measure used to cluster text data. It considers as similarity measure, the cosine of the angle between two non-zero vectors of an inner product space. To compute cosine similarity, the text needs to be converted to a vector representation. While there exists other popular techniques, the current work uses the basic Python implementation which considers term frequency for vector representation.

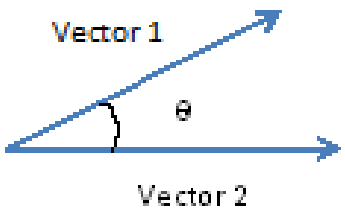

Figure 1: Vector Representation of Text

IV. RESULTS AND DISCUSSION

Detailed experiments were carried out on the reported dataset with the model answers listed in Table 1, and the results tabulated. Considering the fact that the model answers are all single sentences and some student answers are longer, two schemes of evaluation are considered. In scheme 1, the individual Cosine Similarity measures of every sentence of student answer with every sentence of model answer are summed. In scheme 2, the highest Cosine Similarity is considered. The value is then normalised w.r.t to the number of sentences in MA and SA. Table 3 shows the performance of the proposed methodologies for Question No. 1. Pearson Correlation Coefficients is used as a measure of correlation between average score of two human evaluators and Cosine Similarity score (Scheme 1 and Scheme 2) of student answer and model answer,for each the question.

Table 3: System Performance for Question No. 1

Student No.	Eval 1	Eval 2	Avg (Eval 1, Eval 2)	Score (Scheme 1)	Score (Scheme 2)
1	1.6	1.5	1.55	1.13	1.5
2	0.8	0.5	0.65	0.94	0.94
3	1.6	0.5	1.05	0.88	0.88
4	1.6	1.5	1.55	0.66	1.26
5	1.6	0.5	1.05	0.54	0.54
6	1.6	0.5	1.05	1.24	1.24
7	1.6	0.5	1.05	0.5	0.5
8	0.8	1.5	1.15	1.2	1
9	1.6	0.25	0.925	0.74	1.48
10	1.6	1.6	1.6	0.8	1.26
11	1.6	1.5	1.55	1.1	1.1
12	1.6	0.2	0.9	1.06	1.06
13	1.6	1	1.3	1.16	1.16
14	1.6	0.5	1.05	1.02	1.54
15	0.4	0.1	0.25	0.8	0.8
16	1.6	1	1.3	1.16	1.16
17	0.8	0.5	0.65	0.8	0.8

Figures 2, 3 and 4 shows the plots for Questions 1, 2 and 3, while Table 4 shows for different schemes of evaluation.

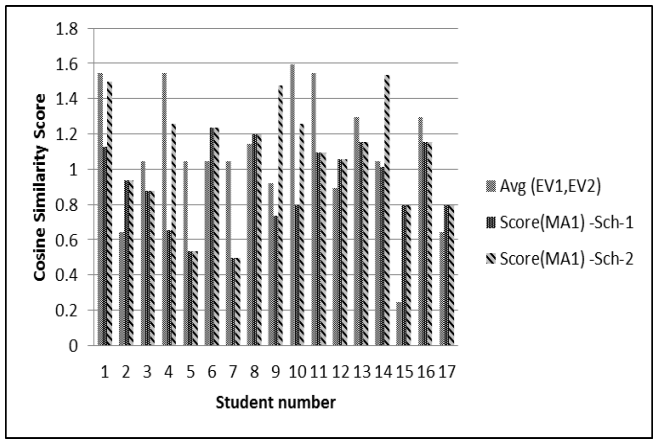

Figure 2: Plot showing performance on Question 1

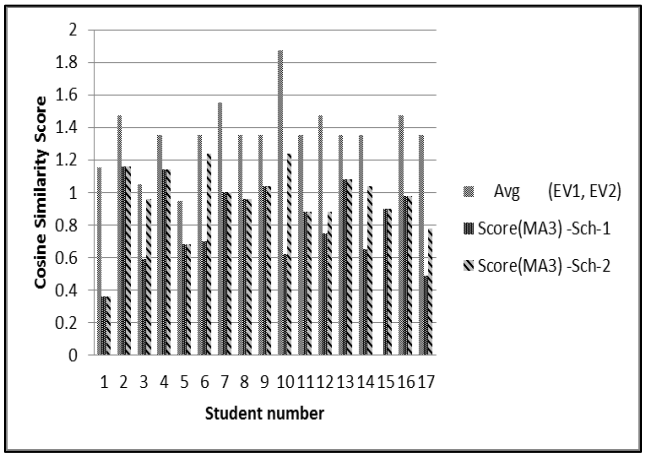

Figure 3: Plot showing performance on Question 2

Considering that every answer is a unique vector that carries some information, these are compared against the standard answers. Accuracy of the students' response can thereafter be measured based on the cosine similarity and the correlation returned by evaluated values.

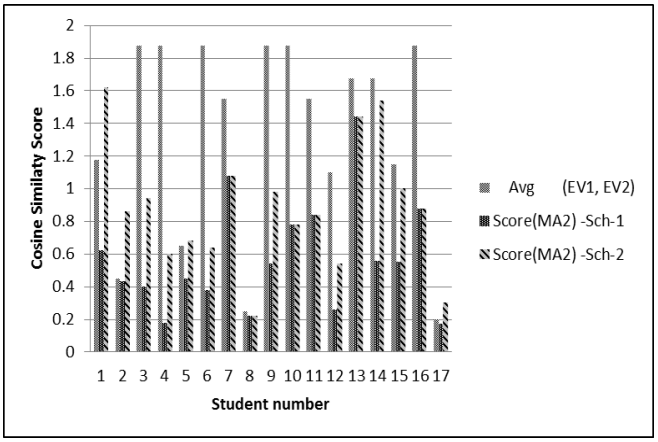

Figure 4: Plot showing performance on Question 3

The correlation shows that the vector similarity measure of student answer with model answer nears the human evaluators. Measured against the average of two human evaluators the values returned show sufficient statistical significance for Scheme 2 of scoring as discussed above.

Table 4: Pearson Correlation Coefficients

	Q1		Q2		Q3	
	Sch1	**Sch2**	**Sch1**	**Sch1**	**Sch2**	**Sch1**
Avg (EV1, EV2)	**0.2052**	**0.4457**	**0.4952**	**0.2052**	**0.4457**	**0.4952**

V. CONCLUSION AND FUTURE SCOPE

Text processing is a complex task. It becomes even more difficult when the knowledge content is under consideration. The lack of a proper, well accepted measure being the primary issue, the basic structural issues of natural languages also play a role. The reduced size of text in this particular case also results in lack of information extraction possibilities.

Nevertheless, the proposed simplistic method does reasonably well considering that it uses the most basic functions and deals with very short answers. The second scheme shows reasonably acceptable correlation with the average score of two human evaluators. The measures, tabulated in Table 4 reasonably substantiate the accuracy of the proposed method.

The proposed method may further be improved through the use of other embedding's as the term frequency model also performs significantly better with larger corpuses which forms part of future work.

REFERENCES

[1]. Madhumita Chakraborty, 2018, 'Here's why DU teachers are not evaluating answer papers since May 24', *Hindustan Times*, June 15, 2018.

[2]. K. A. Gafoor, T.K. Umer Farooque, "Incongruence in Scoring Practices of Answer Scripts and Their Implications: Need for Urgent Examination Reforms in Secondary Pre-Service Teacher Education", Proceedings of UGC sponsored national seminar on Fostering 21st Century Skills: Challenges to Teacher quality, August 22-23, 2014, Kerala, pp. 2-5.

[3]. Ritu Sharma, 2017, ''Model Rules': Board to train teachers how to evaluate answer-sheets', *The Indian Express*, September 8, 2017.

[4]. Priyanka Dhondi, 2015, 'Different Types of Questions in E-learning Assessments', *ElearningDesign*, CommLabIndia, January 20, 2015.

[5]. Komi Reddy Deepika, 2014, 'Different Types of Assessments Used in E-learning', *ElearningDesign*, CommLabIndia, June 27, 2014.

[6]. S. Ramesh,"Exploring the potential of Multiple Choice Questions in Computer Based Assessment of Student Learning", *Malaysian Online Journal of Instructional Science*, 2005, Vol. 2.

[7]. M. Bush, "Alternative Marking Schemes Fof On-line Multiple-choice Tests", Proceedings of 7th Annual Conference on the Teaching of Computing, Belfast, 1999.
[8]. Megan Clendenon, Hannah Holley, Mauro Schimf 'Responding to Short Answer and Essay Questions', StudentCaffe, Updated on April 2018.
[9]. Allen Grove, 'What is the ideal word count for the short answer on the common application?', ThoughtCo, Updated on 22 November, 2018.
[10]. S. Burrows, I. Gurevych, B. Stein, "The Eras and Trends of Automatic Short Answer Grading", *International Journal of Artificial Intelligence in Education* , 25, IOS Press, pp. 60-117, 2015.
[11]. Y. Li, A. Tripathi, A. Srinivasan, "Challenges in Short Text Classification: The Case of Online Auction Disclosure", Tenth Mediterranean Conference on Information Systems (MCIS), Paphos, Cyprus, September 2016.
[12]. M. Hermet, S. Szpakowicz , L. Duquette and S. N. Leuven, "Automated Analysis of Students' Free-text Answers for Computer-Assisted Assessment", *Proceedings of TAL and ALAO Workshop*, pp. 835--845, 2006.
[13]. A. Adam, A. Ismail, A. Rafiu, A. Mohamed, G. Shafeeu, M. Ashir, "Pedagogy and Assessment Guide", National Institute of Education, Male, Maldives, 2014. Accessed on: 02nd September 2018.
[14]. J. Burstein, R. Kaplan, S. Wolff, & C. Lu, Using Lexical Semantic Techniques to Classify Free-Responses. In E. Viegas, editor, Proceedings of the ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons, pages 20–29, Santa Cruz, California. Association for Computational Linguistics, 1996.
[15]. J. Cowie, Y. Wilks, Information Extraction. In R. Dale, H. Moisl, and H. Somers, editors, Handbook of Natural Language Processing, chapter 10, pages 241–260. Marcel Dekker, New York City, New York, First Edition, 2000.
[16]. C. Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries. In M.-F. Moens and S. Szpakowicz, editors, Proceedings of the First Text Summarization Branches Out Workshop at ACL, pages 74–81, Barcelona, Spain. Association for Computational Linguistics, 2004.
[17]. M. Mohler, R. Mihalcea, Text-to-text Semantic Similarity for Automatic Short Answer Grading. In A. Lascarides, C. Gardent, and J. Nivre, editors, Proceedings of the Twelfth Conference of the European Chapter of the Association for Computational Linguistics, pages 567–575, Athens, Greece. Association for Computational Linguistics, 2009.
[18]. Automated Short Answer Grading – Dataset 1 [https://sites.google.com/site/uditkc/home/reading-stuff]

AUTHORS PROFILE

Mrs. C. Chakraborty, currently employed with the department of Information Technology, Sikkim Manipal Institute of Technology, Sikkim Manipal University, has over 18 years teaching experience in various universities across India. Currently pursuing her doctoral research from the University of Science & Technology, Meghalaya, she is actively engaged in research on Education Technology. Her teaching interests lie in the fields of Soft Computing, Speech Processing and Deep Learning.

Dr. Udit Kr. Chakraborty, is currently employed with the Sikkim Manipal Institute of Technology, Sikkim Manipal University, as an Associate Professor in the department of Computer Science & Engineering. He has almost 20 years combined industry and academic experience and has published a good number of research articles in journals of international repute. Dr. Chakraborty has also co-authored a book on Soft Computing, published by Pearson and features in the Editorial Board of some noted international journals. His research interest lies in the fields of Algorithms, Artificial Neural Networks and Education Technology.

Dr. Bhairab Sarma, is working as an Associate Professor, with the University of Science & Technology, Meghalaya. His research area is Natural Language Processing. He has published many research papers in the field of Database, Data Mining, Artificial Intelligence, Computational Linguistics and Modeling & Simulation.

