Automatic Short Answer Grading Using Universal Sentence Encoder

Chandralika Chakraborty[1], Rohan Sethi[2], Vidushi Chauhan[2], Bhairab Sarma[3], and Udit Kumar Chakraborty[1](✉)

[1] Sikkim Manipal Institute of Technology, Sikkim Manipal University, Sikkim, India
udit.kc@gmail.com
[2] Dell Technologies, Bangalore, India
[3] University of Science & Technology Meghalaya, Baridua, India

Abstract. Automatic Evaluation of Text Answers, popularly known as Automatic Short Answer Grading (ASAG) is an area of research and development currently. The widespread acceptance of online learning and increased number of enrolments in such courses has necessitated the creation of a method that can be applied across platforms for all types of supply based answers. The current paper proposes a simple technique using the Deep Learning Based Universal Sentence Encoder to generate vectors for each answer. These vectors can then be compared against vectors generated from model answers to get the final score for the student's answer. Experimental results show that for a sizeable dataset, the approach works well and can be considered a reliable approach.

Keywords: Automatic Short Answer Grading · Vector · Word Embedding · Cosine Similarity · Universal Sentence Encoder · Confidence

1 Introduction

Examinations are an important component of the teaching learning process. The basic purpose being reinforcement of the learning accomplished by the student. Evaluation of learner's responses to questions therefore has to be correct, uniform and impartial. However, these criteria are not always met. In schools and universities, with a large number of enrolments, the teacher–student ratio can be high. In the Indian scenario, this ratio is as high as 1:20 in technical courses [1], resulting in even the time allotted for evaluation being insufficient [2]. The issue gets further complicated by the introduction of Massive Open Online Course (MOOC) based content delivery with universities pitching in with online lecture delivery resulting in increased enrolments.

Manual evaluation of a large number of answer scripts has various problems associated with it. While timely delivery is the most highlighted, uniformity in evaluation standards is also an issue. Evaluation styles differ, and so do expectations from learners. Uniformity may be achieved to a certain degree using rubrics, but the basic interpretation is still human dependent. Additionally, rubrics preparation needs time and training [3].

Automatic short answer grading is the task of assessing short natural language responses to objective questions using computational methods [4]. These usually cover

M. E. Auer et al. (Eds.): ICL 2022, LNNS 633, pp. 511–518, 2023.
https://doi.org/10.1007/978-3-031-26876-2_49

answers to questions with multiple choices, fill-in-the missing-words, single sentence answers and short answers written in natural languages. The computational complexities involved in evaluating the natural language answers have resulted in wide acceptance of multiple choice and single word answer-based questions. Their popularity being credited to objectivity and quantifiability, the popular segment has their drawbacks, as it is difficult to check the learners' knowledge and understanding of the proof and the theoretical aspects [5]. Apart from these, the close-ended questions can also be scored from by using guess work [6].

On the other hand, open-ended questions expect the learner to construct answers in natural languages using knowledge, logic and linguistic abilities simultaneously. These answers have no fixed methods and may also have multiple solutions [7]. This contributes substantially towards the difficulties in strategizing and solution building for automated evaluation of text-based answers.

To standardize ASAG, short answers have been defined to meet the following five-point criteria, viewed from the perspective of evaluation have to meet the five-point criteria [4]:

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.

ASAG returns substantial literature consisting of different approaches tried by esteemed researchers. However, a solution acceptable to all has not yet been formulated. A comprehensive survey of Automatic Short Answer Grading is available in the highly cited work by Burrows et al. [4], which covers in much detail the progress made till 2015. Recent approaches to ASAG show trends shifting gradually to Machine Learning and Deep Learning based techniques. This approach needs the text information to be converted to vector representations for use in Artificial Neural Network (ANN) models. There are various embedding techniques available for the purpose. However, not all suit every requirement and the choice parameters are yet to be correctly identified. This paper reports experiments conducted on text data for ASAG using Universal Sentence Encoder. The results are presented with discussions on the suitability of the encoder on the data used with analysis.

2 Literature Review

ASAG has been an area of active research for quite some time now. The paper by Burrows et al. [4] provides a comprehensive survey of the developments made prior to 2015. Post 2015, the focus has mostly been directed towards Artificial Intelligence and Machine Learning based approaches. Supervised machine Learning approaches use well labelled data to train networks in performing tasks on data that it has not been exposed

to before. Employed to ASAG, unsupervised techniques use word and text similarity measures to grade student's responses with respect to reference answers [8]. In his thesis, Roy [8] proposes a number of novel ASAG techniques based on machine learning and computational linguistics principles with extensive empirical evidence on multiple datasets. The work reports about 20% improvement over existing results on correlation. This also presents brief successes of about 8% improvement using ensemble based techniques. The primary findings of this work however point towards use of ensembles for better performance.

Most work using Deep Learning use embedding techniques for vectorization and find the vector distances between some standard identified for the problem. In [9], the authors used a bag of words and k-means clustering to group similar answers. Although the paper reports a correlation of 0.83, the sample space is rather small with only 29 students' responses from ten assignments and two examinations. Further, the approach did not work for synonyms used in the students' responses.

Similar work, done by Lubis et.al. [10] tried using word embedding with only one model answer and semantic analysis. The experiments were conducted on a rather limited dataset of 224 responses with a correlation value of 0.7085. The essence of word embedding consisting of capture of the context could not be explored here as semantic analysis measures were taken separately.

The actual process of evaluation being dependent on the input embedding, a proper evaluation of all embedding models needs to be conducted. To this end, the work by Ghavidel et.al. [11], compares Bidirectional Encoder Representations (BERT) an autoencoder with XLNET, a bidirectional transformer for performance. However, the dataset being rather limited, a reliable comparison does not come out. On the chosen dataset, both perform equally well.

The work done, as presented in this paper, tries to mitigate the complexity of the ASAG process, by adopting the classical approach to learners' response evaluation. Using the Universal Sentence Encoder over other task specific embedding techniques, the experiments were conducted on a reasonably sized dataset having 1272 student's answers. To ensure correctness of the process, the scores were compared against the average score of five human evaluators. The results show substantial improvement over currently reported results.

3 ASAG Methodology

While multiple techniques have been employed over the years to automatically grade or score short answers, none has been accepted as the gold standard for ASAG. As against the information extraction [12] or student's wisdom approach [13] the current paper sticks to the classical approach. In this work, the student's answers are evaluated with respect to model answers which are expected to be provided by subject experts. The methodology is implemented as shown in Algo BasicASAG.

Algo BasicASAG

1. Begin
2. For each of the Model Answers, do
    a. Encode $ModelAnswer_i$
    b. Store Vector as MA_i
3. For each of the Learners' Responses
    a. Encode $LearnerResponse_i$
    b. Store Vector as LR_i
    c. For each vector MA_j
        i. Compute Cosine Similarity between LR_i and MA_j
        ii. Store value as CS_{ij}
    d. Compute $Score_i = \frac{\sum_{j=1}^{n} CS_{ij}}{n}$
4. End

The scores are actually average of the cosine similarity values between the vectors. Cosine similarity is a measure of similarity that is used to compare documents represented as vectors. If x and y are two vectors for comparison, the cosine similarity is computed as:

$$sim(x, y) = \frac{x.y}{||x||||y||} \tag{1}$$

In Eq. 1, ||x|| is the Euclidean Norm of vector x, and is calculated as:

$$\sqrt{x_1^2 + x_2^2 + x_3^2 + ... + x_n^2} \tag{2}$$

The measure returns the cosine of the angles between the two vectors under consideration. A cosine value of 0 signifies orthogonal vectors, meaning thereby that the vectors are dissimilar. A cosine value of 1 signifies that the vectors are identical.

While Euclidean distance returns the distance between two points, a cosine measure is better representative of concepts as embedded in texts. Euclidean distances may vary even due to document sizes, but cosine measure is more advantageous when the purpose is to measure the document's perspective [14].

4 Universal Sentence Encoder

Word embedding or encoding is used in Artificial Neural Networks. The purpose is primarily to encode linguistic information to numerical data that can be handled by the neural model. Many embedding techniques are available for the purpose, some popular ones being Word2Vec, BERT, FastText etc. The present work used Universal Sentence Encoder to encode the sentences into vectors.

A major drawback of other encoding techniques lies in the way they function. These methods encode individual words separately and then encode the sentence by appending

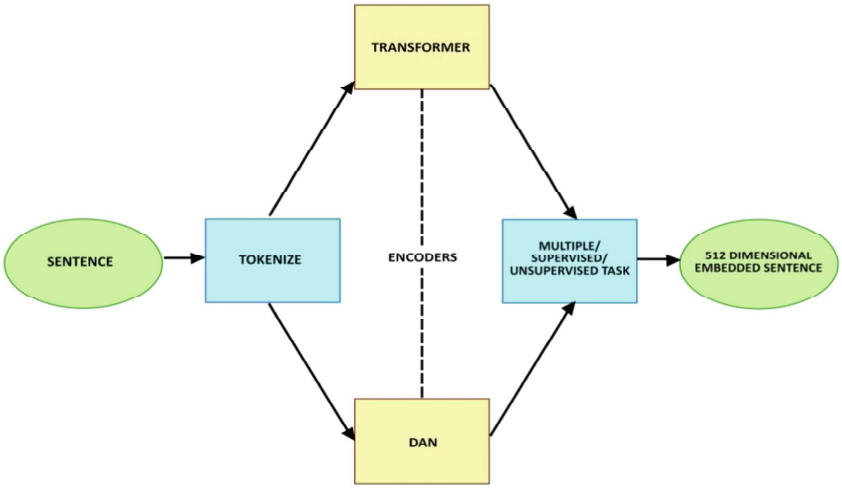

Fig. 1. Universal Sentence Encoder process

the words vectors. As a result there is information loss and word order information fails to be embedded in the vector.

The Universal Sentence Encoder encodes text into high-dimensional vectors that can be used for text classification, semantic similarity, clustering, and other natural language tasks. Implemented in two models, as shown in Fig. 1., namely the Transformer Encoder and the Deep Averaging Network (DAN), the encoder returns 512-dimensional numeric vector representations of each sentence or even a short paragraph. The current paper uses the DAN model as it is less computation-intensive with marginal loss of accuracy [15].

5 Experimentation and Results

The method described in Sect. 3, *Algo BasicASAG*, was implemented in Python and executed on the Kaggle dataset. The dataset consisted of 1272 answers from which five (05) full-scoring answers were identified as model answers. These model answers were considered as benchmarks for the correctness and the other responses were compared.

Initially, the five model answers were encoded using the Universal Encoder. Subsequently, for each of the student's responses, the answers were encoded using the Universal Sentence Encoder and the vectors generated compared with each of the five model vectors for cosine similarity.

The code executed on Google Collab on 1272 number of answers, compared with the rounded average of score of 5 model answers returned results as shown in Fig. 2. These results are summarized and presented in Table 1 and plotted in Fig. 3.

The numbers show that 38.8% of answers have exactly the same evaluation scores returned by the automated evaluation scheme and 53% had a deviation by 1 mark. If benevolence is considered as a factor for evaluation, then the instances where the

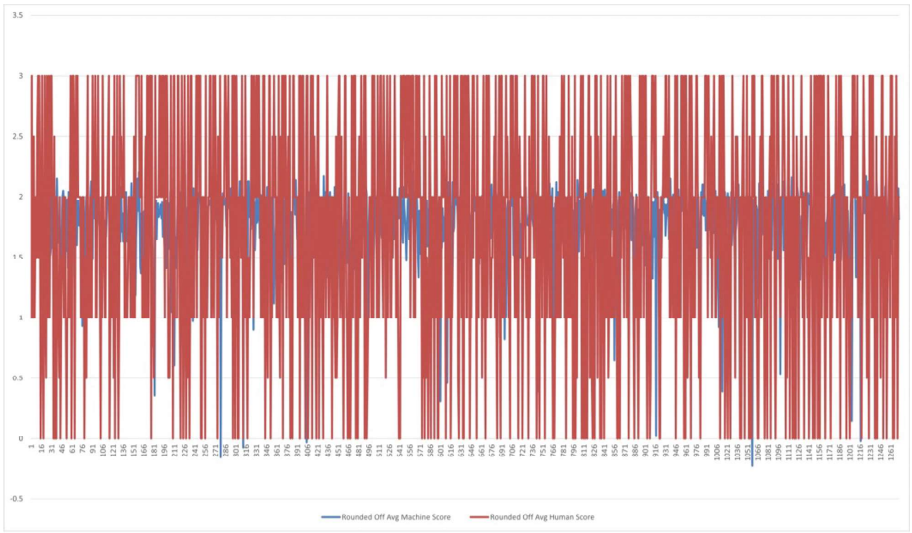

Fig. 2. Plot of Rounded Average Score

Table 1. Summary of Results

xxxxx	No Diff	Higher by 1 mark	Higher by 2 marks	Higher by 3 marks
Rounded Avg. of Hum. Sc	494	308	73	0
Rounded Avg. of Mac. Sc	494	378	17	2

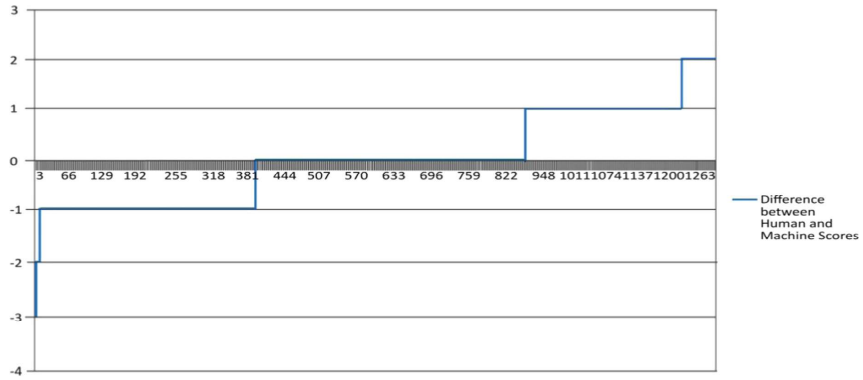

Fig. 3. Comparison Chart

machine has graded the answer higher than the human evaluators by 1 mark may also be considered correct and the accuracy of the approach would be up by 30% to 68.55%.

Considering the task as proving the null hypothesis, as in Eq. 3 and the alternative hypothesis shown in Eq. 4, a both-tail z-test was carried out:

$$H_0 : \mu_d = 0 \tag{3}$$

$$H_1 : \mu_d \neq 0 \tag{4}$$

The details are shown through equations Eq. 5 to.
Checking for 95% confidence:

$$P(-z_{\alpha/2} \leq z \leq z_{\alpha/2} = 0.95 \tag{5}$$

$$P(z > z_{\alpha/2}) = 0.025 \tag{6}$$

Computing the mean difference between the average scores of human evaluators and the automatic evaluation scheme proposed, and the standard deviation, the values are found to be 0.034 and 0.9169 respectively.

$$\mu_d = 0.034 \tag{7}$$

$$\sigma = 0.9169 \tag{8}$$

Computing the z-score using Eq. 9, the value obtained is 1.32:

$$z = \frac{\mu_d - 0}{\sigma/\sqrt{n}} \tag{9}$$

Considering the confidence level of 0.95, as shown in Eq. 5, the z-score for 0.025, as shown in Eq. 6, is 1.96.

Therefore, as the value 1.32 lies within the interval − 1.92 to 1.92, it can be said that the null hypothesis of Eq. 3 holds good.

6 Conclusion

The proposed method for Automatic Short Answer Grading is simplistic in approach and effective in computing the scores. The vectors being considered reflective of the knowledge content of the answers, a cosine similarity between two such vectors would return the difference in score. However, the method is reliant on the efficacy of the Universal Sentence Encoder and it remains to be seen whether it works with same or at least similar accuracy with other types of answers. The deviation of 2 or more marks in a 3 marks question, though limited to 6.4% is also an area of improvement where some effort is due.

References

1. Chakraborty, M.: Here's why DU teachers are not evaluating answer papers since May 24. Hindustan Times, June 15, 2018 (2018)
2. Gafoor, K.A., Umer Farooque, T.K.: Incongruence in scoring practices of answer scripts and their implications: need for urgent examination reforms in secondary pre-service teacher education. In: Proceedings of UGC sponsored national seminar on Fostering 21 st Century Skills: Challenges to Teacher quality, August 22–23, 2014, Kerala, pp. 2–5 (2014)
3. Sharma, R.: (2017), "Model Rules: Board to train teachers how to evaluate answer-sheets". The Indian Express, September 8, 2017
4. Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. Int. J. Artif. Intell. Educ. **25**(1), 60–117 (2014). https://doi.org/10.1007/s40593-014-0026-8
5. Chang, S.-H., Lin, P.-C., Lin, Z.C.: Measures of partial knowledge and unexpected responses in multiple-choice tests. Educ. Technol. Soc. **10**(4), 95–109 (2007)
6. Lau, P.N.K., Lau, S.H., Hong, K.S., Usop, H.: Guessing, partial knowledge, and misconceptions in multiple-choice tests. J. Educ. Technol. Soc. **14**(4), 99–110 (2011). http://www.jstor.org/stable/jeductechsoci.14.4.99
7. Yee, F.P.: Using Short Open Ended Mathematics Questions to Promote Thinking and Understanding. National Institute of Education, Singapore (2002)
8. Roy, S.: New Techniques for Automatic Short Answer Grading [Doctoral thesis, Indian Institute of Science, Bangalore] (2017)
9. Suzen, N., Gorban, A.N., Mirkes, E.M.: Automatic short answer grading and feedback using text mining methods, ArXiv:abs/1807.10543 (2018)
10. Lubis, F.F., et al.: Automated short answer grading using semantic similarity based on word embedding. Int. J. Technol. **12**(3), 571–581 (2021)
11. Ghavidel, H.A., Zouaq, A., Desmarais, M.C.: Using BERT and XLNET for the Automatic Short Answer Grading Task. In: Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020) - Volume 1, pages 58–67 (2020)
12. Hasanah, U., Permanasari, A.E., Kusumawardani, S.S., Pribadi, F.S.: A review of an information extraction technique approach for automatic short answer grading. In: 2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), pp. 192–196 (2016). https://doi.org/10.1109/ICITISEE.2016.7803072
13. Roy, S., Dandapat, S., Nagesh, A., Narahari, Y.: Wisdom of students: a consistent automatic short answer grading technique. In: Proceedings of the 13th Intl. Conference on Natural Language Processing, Varanasi, India. December 2016, pp. 178–187 (2016)
14. Wang, J., Dong, Y.: Measurement of text similarity: a survey. Information **11**(9), 421 (2020). MDPI AG. http://dx.doi.org/https://doi.org/10.3390/info11090421
15. Cer,'D., et al.: Universal Sentence Encoder (2018). arXiv:1803.11175v2 [cs.CL]. https://doi.org/10.48550/arXiv.1803.11175