

Automated Assessment and Grading System for Short Answers

Riddhi Kundal, Bhagvati Parekh

Bhakt Kavi Narsinh Mehta University

Corresponding author: Riddhi Kundal, Email: riddhi.joshi931@gmail.com

Automatic short answer grading is a task that uses computational algorithms to analyse responses written in natural language. Several years ago, there was a computer-aided assessment. This work has been done mostly for the English language in a variety of languages. Despite this, due to the complexity of the Indian natural language, not much study has been done on Indian languages. The goal of this research is to provide a comprehensive overview of ASGS (Automatic Short Answer Grading System). The methods and procedures utilised in the automatic short answer grading system for English and other languages are discussed in this study.

Keywords: ASAG (Automated Short Answer Grading), Natural Language Processing, Gujarati, Sentence Recognition, OCR, HCR, PCR

1 Introduction

The Evaluation Process is a significant part of the education system. Every year millions of students give different kinds of examinations, and teachers evaluate their attempts. The evaluation process is a very lengthy and responsible task for evaluators. Government and private institutions are also paying evaluators. As day by day, the number of student's increases, so evaluation is becoming more complex and tedious. Sometimes manual assessment is becoming biased too. So, to overcome this tedious task and make the review smooth and unbiased, computer- aided grading methods have been generated, called Automated Assessment or Automated Evaluation system.

There are various different forms of questions through which evaluators can judge the knowledge of the students like multiple choice questions, Essay Type questions, Subjective answers etc. Following Fig. 1 shows the types of questions.

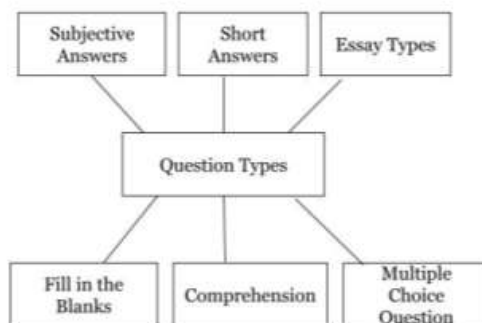


Figure 1 Types of Questions

In India there are 1.43 crores students who give an annual exam every year for various boards for schooling. Throughout the year students are also evaluated through different class tests. Multiple choice questions are the main choice of teachers for evaluating the students as it is a very easy method.

In the recent COVID situation many institutes conducted online examinations. For that they have chosen a multiple-choice answer system. Multiple choice questions are also easy for the students to mug up, but evaluators cannot judge the knowledge about the concept. So, Short answers are the option for evaluators to judge the students in the online examination system.

In the mechanism of the automated evaluation system the tedious process of evaluating the answers has been removed. Because manual work has been reduced the chances of mistakes and biases has also been reduced. After the declaration of board results, many students apply for rechecking and mistakes have also been found. so in these cases automated evaluation is a boon. The research in this field is ongoing from the early '70s to today. There are so many techniques available for different purposes, and all have their benefits. Like consistent performance for large class size, cost effective and also provide a formalisation of assignment criteria. Automatic grading systems can provide immediate responses wherein the test taker would need to wait for human assessors to finish the grading on manual systems.

2 Challenges

As there are multiple ways for student marking like MCQ, Comprehensive answers, Essay type answers, short one or two liner answers etc., which method is used for marking the student is important. If the method used like essay type questions or comprehensive or short type questions, then recognition of student responses plays a vital role. In printed format it is a little bit easy to recognize the responses as there are many APIs available which can identify printed characters with great accuracy. If student responses can be taken in handwritten format, then recognition is a very big challenge.

Fig. 2 shows the methods of student response.

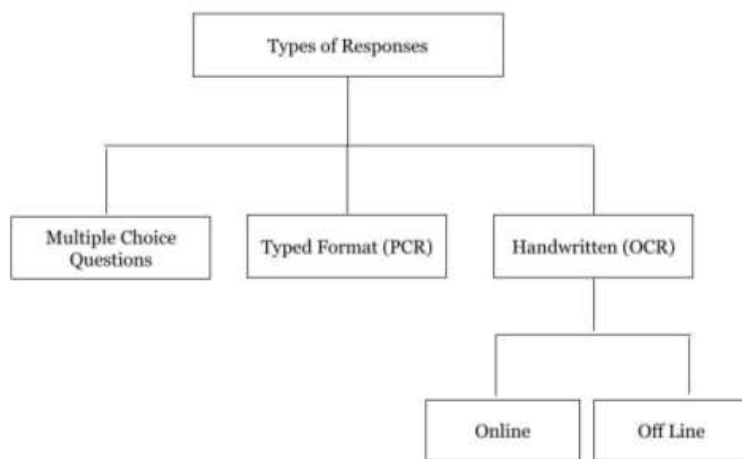


Fig. 2 - Types of student Responses

Before recognising the text, it must be pre-processed. First the paper on which student responses have taken needs to be scanned. It will give you the grey scale image. That image first converts into black and white called binarization. There are different strokes in handwritten text, so it must be managed so thinning and skeletonization is needed. After noise removal the text must be segment like line segmentation and word segmentation. After this feature of the given text have to be extracted. Feature extraction and classification both the steps are very important. There are various techniques through which features can be extracted such as, power & parabola arc fitting, diagonal feature extraction, transition feature extraction etc. Classification is used for pattern matching. There are various classifiers available in machine learning and deep learning. Classification is basically the phase where decision making is done. Here the features are used that we get from feature extraction. Different classifiers can be used for the purpose of recognition like K-NN (K nearest neighbour, SVM (Support Vector Machine), Artificial Neural Network (ANN) etc. Following figure Fig. 3 shows the pre-processing steps of HTR.

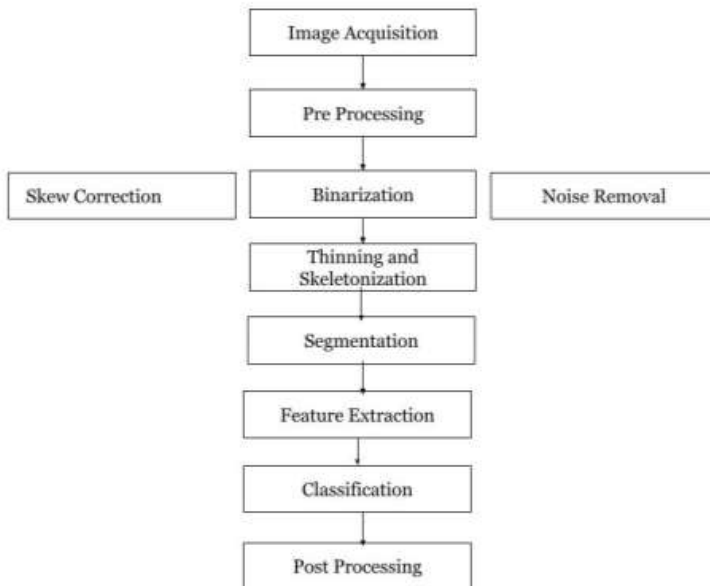


Fig. 3 Steps for HTR

3. Review of Existing system

There are multiple Auto scoring systems available for different languages. The review defines the system for different systems which has been developed for short or essay type answers. Like P.Selvi and Dr.A.K.Bnerjee (2010) have proposed ontology, semantic similarity, and statistical methods for short answer grading systems. The authors achieved 59% accuracy concerning traditional vector-based methods. [1]

Michael A.G. Mohler, Razvan Bunescu, Rada Mihalcea (2011) have applied machine learning corpus-based and knowledge-based algorithms on a data set of computer science student's answers. They have achieved good accuracy. [2]

Wael Hassan Gomaa, Aly Aly Fahmy (2014) proposed string-based similarity measures to detect similarities between student and model responses. They also provide immediate feedback to students about their answers. Authors have applied these string-based and corpus-based measures on Arabic short answers. [3]

Dimitrios Alikaniotis, Helen Yannakoudakis, Marek Rei(2016) In this paper, authors have proposed their system using deep neural networks. Researchers have used a data set of essays written by the students of grade 7 to grade 10 of 150 to 550 words. One of the drawbacks is that their system cannot distinguish between the words that appear multiple times. [4]

Udit Kr. Chakraborty and Debanjan Konar, Samir Roy, Sankhayan Choudhury (2016) have proposed a system that will match the short answers which contain only a single sentence. They have done this with the use of clustering techniques. They applied keyword matching from the student's textual responses. The system generates clusters based on keywords. [5]

JCS Kadupitiya, Surangika Ranathunga, Gihan Dias (2016). In this paper, the authors have proposed a hybrid methodology that contains corpus-based and knowledge-based methods. They have applied these techniques to the Sinhala language, which follows free word order. As Sinhala's linguistic resources are not that much rich, authors cannot get great accuracy. [6]

Hebah Rababah, Ahmad T. Al-Taani (2017). The authors proposed a system on Arabic short questions. The authors used cosine similarity measures. The research includes word roots as well synonyms of student answers for accurate results. [7]

Feddy Setio Pribadi, Teguh Bharata Adj, Adhistya Erna Permanasari, Anggraini Mulwinda, and Aryo Baskoro Utomo (2017) analyze several methods that apply the overlapping techniques to determine the degree of similarity between the model responses and students' responses. The research result displays that the method Cosine

Coefficient has better results than the Dice and Jaccard Coefficient methods. The authors have achieved 80% accuracy. [8]

Akeem Olowolayemo, Santhy David Nawi, Teddy Montoro (2018) proposed a system applied to fill in the blanks type of questions for English grammar. This system used Levenshtein distance (LD) and Cosine similarity measures. They have achieved 92% accuracy for LD and 94% accuracy for cosine similarity. [9]

Adidah Lajis, Haidawati Nasir, Normaziah A. Aziz (2019) This research has been done using Natural Language Processing and Convolutional Neural networks. The authors applied these techniques to engineering, Information technology, and management students' answer scripts, achieving excellent results. [10]

Sonakshi Vij, Devendra Tayal, Amita Jain (2019). In this paper, authors applied fuzzy wordnet-based graph methods on student answer sheets for a short answer grading system. This result is applied to the synthetic dataset and has achieved 52.98% accuracy. [11]

Stefano Menini, Sara Tonelli, Giovanni De Gasperis, Pierpaolo Vittorini (2019) used Natural Language Processing and sentence embedding. The work has been done based on distance-based features. The authors applied the technique to statistical exam data of students, and they have achieved 87% accuracy. [12]

Anak Agung Putri Ratna, Lea Santiar, Ihsan Ibrahim, Prima Dewi Purnamasari, Dyah Lalita Luhurkinanti, Adisa Larasat (2020) This system is based on Japanese language. The authors applied the two algorithms are the Singular Value Decomposition (SVD) and the windowing algorithm. The system got 87.78% accuracy. [13]

Neslihan Suzen, Alexander N. Gorban, Jeremy Levesley, Evgeny M. Mirkesa (2020). This research has been applied on the UK's GSEC exam and provides feedback to the students. Authors first use standard data mining techniques on student answers corpus with model answers. Then they have found similarity measures based on the number of common words. For this, they have applied a clustering algorithm. [14]

4. Conclusion and Future Directions

In this paper, we have done a survey of research being done towards the automated scoring system. Most of the paper focuses on different Machine learning techniques for classification and scoring. It is observed that Statistical methods and corpus-based methods have proved to be state of the art for automated scoring. Recent Deep learning techniques like ANN, CNN, RNN have outperformed traditional techniques for the classification of data.

We have identified the future scope of research in the area of automated scoring systems and possible tasks that may be needed like creation of corpus for Indian languages for different types of question answers. Generally, students' responses are taken on answer sheets like papers so handwritten recognition and matching work needs to be researched more.

References

- [1] Selvi, P. and Banerjee, A.K., 2010. The automatic short-answer grading system (ASAGS). arXiv preprint arXiv:1011.1742.
- [2] Mohler, M., Bunescu, R. and Mihalcea, R., 2011, June. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (pp. 752-762).
- [3] Gomaa, W.H. and Fahmy, A.A., 2014. "Arabic short answer scoring with effective feedback for students." International Journal of Computer Applications, 86(2).
- [4] Alikaniotis, D., Yannakoudakis, H. and Rei, M., 2016. Automatic text scoring using neural networks. arXiv preprint arXiv:1606.04289.

- [5] Chakraborty, U.K., Konar, D., Roy, S. and Choudhury, S., 2016. Intelligent fuzzy spelling evaluator for e-Learning systems. *Education and Information Technologies*, 21(1), pp.171-184.
- [6] Kadupitiya, J.C.S., Ranathunga, S. and Dias, G., 2016. Sinhala Short Sentence Similarity Measures for Short Answer Evaluation based on Corpus-Based Similarity.
- [7] Rababah, H. and Al-Taani, A.T., 2017, May. An automated scoring approach for Arabic short answers essays questions. In 2017 8th International Conference on Information Technology (ICIT) (pp. 697- 702). IEEE.
- [8] Pribadi, F.S., Adjil, T.B., Permanasari, A.E., Mulwinda, A. and Utomo, A.B., 2017, March. Automatic short answer scoring using word overlapping methods. In AIP Conference Proceedings (Vol. 1818, No. 1, p. 020042). AIP Publishing LLC.
- [9] Olowolayemo, A., Nawi, S.D. and Mantoro, T., 2018, September. Short Answer Scoring in English Grammar Using Text Similarity Measurement. In 2018 International Conference on Computing, Engineering, and Design (ICCED) (pp. 131-136). IEEE
- [10] Lajis, A., Nasir, H. and Aziz, N.A., 2019, August. NCI Evaluation: Assessment of Higher Order Thinking Skills via Short Free Text Answer. In 2019 IEEE International Conference on Smart Instrumentation, Measurement and Application (ICSIMA) (pp. 1-5). IEEE.
- [11] Vij, S., Tayal, D. and Jain, A., 2020. A machine learning approach for automated evaluation of short answers using text similarity based on WordNet graphs. *Wireless Personal Communications*, 111(2), pp.1271-1282.
- [12] Menini, S., Tonelli, S., De Gasperis, G. and Vittorini, P., 2019, November. Automated Short Answer Grading: A Simple Solution for a Difficult Task. In *CLiC*.
- [13] Ratna, A.A.P., Santiar, L., Ibrahim, I., Purnamasari, P.D., Luhurkinanti, D.L. and Larasati, A., 2019, October. Latent Semantic Analysis and Winnowing Algorithm Based Automatic Japanese Short Essay Answer Grading System Comparative Performance. In 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST) (pp. 1-7). IEEE.
- [14] Stizen, N., Gorban, A.N., Levesley, J. and Mirkes, E.M., 2020. Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science*, 169, pp.726-743.