# Automatic Grading For Short Answer Script (AGSAS)

**Navpreet Singh Aulakh[1], Mandar Mhatre[2], Rishabh Gupta[3], Rahul Reddy[4], Prof. Sunil Shelke[5]**
[1, 2, 3, 4, 5] Dept of Computer Science and Engineering
[1, 2, 3, 4, 5] Pillai College of Engineering, Navi Mumbai, India - 410206

***Abstract-*** *During checking of answers, teachers usually go through the same lengthy process. This repetitive and monotonous process can cause errors in grading. The same problem can be solved by automatic grading techniques which use Natural Language Processing to grade short answers without the need of any external assistance. This is applicable to technical courses, such automated grading techniques will make grading papers not only easy and quick but ethical as well. The grader takes the short answers and divides it into parts or points and checks them individually and then makes sense of all of them as a whole answer to then finally come to its conclusion where it assigns the marks to the respective answer. Some of the factors which the grader takes into account are any technical or important terms in the answer as well as the overall depth and meaning of it. Techniques used in this domain can be broadly classified into two types based on their requirement of reference or model answers, Our work focuses on both of these types, i.e questions having a model answer and questions without a model answer. Limitation of these existing techniques is that they either rely on the answers written by students too much or require a well labelled dataset with a variety of scores. So combining the two techniques and using weighted combinations of scores produced by them is the  proposed technique used in this work.*

***Keywords***- Natural Language Processing, Automatic Grading, Weighted Combination

## I. INTRODUCTION

Automatic short answer grading (ASAG), also known as short answer scoring (SAS), is the task of grading students' short responses, constructed in natural language, with respect to instructor-provided reference answer(s) and/or scoring schemes.

**Length**: The word short in “short answer” has lacked precise definition in the literature. We find following versions in various text:

a) “from about one phrase (several words) up to one paragraph”.
b) “phrases to three to four sentences”.
c) “a few words to approximately 100 words”.

While the above definitions are imprecise, it is typically not difficult to decide whether an answer can be considered as short. Intuitively, they are not as long as essays but long enough for students to express the answers. The Automatic Grading for Short Answer Script (AGSAS), aims to grade the answers of the students more optimally then the previous techniques. It does so by combining the existing methods/techniques.

The objectives of this work are as follows:

a) To study the ASAG techniques and identify their limitations that may help to suggest an optimal approach which may overcome the drawbacks of existing methods.
b) To understand methods of grading for grading systems and combining the methods that may help the system to give the proper and more accurate grades to the answer.
c) To identify evaluation metrics used for performance analysis of the grading system.

## II. LITERATURE SURVEY

***A. Wisdom of Students: A Consistent Automatic Short Answer Grading Technique[4]:***This work proposes an unsupervised technique of grading answers without requiring a model answer.. It consists of 2 steps:

Step 1] : Sequential pattern mining problem - for matching the answers by finding the pattern. This basically means that it tries to create its own answer model answer based on the answers written by the students
Step 2] : Intuition Driven hypothesis - used for grading i.e. it matches the sequential pattern found in the previous step with all the answers and grades them accordingly

Many ASAG techniques have fluctuations in the grading of answers, this Intuition Driven technique shows no such fluctuation. This work provides significantly better correlation than all word similarity based ASAG techniques. This technique would perform the best when all answers are correct in the same manner and worst if in an unlikely case all are wrong in the same manner.

***B. Distributed Vector Representations for Unsupervised Automatic Short Answer Grading[1]:***This paper proposes an unsupervised technique of grading answers The technique consists of 2 steps:

Step 1] : Providing model answer and student answer as datasets in which some datasets have presence of a weighted scoring scheme for each question, which demonstrates promise in improving unsupervised ASAG performance when used.

Step 2] : It is observed that for ASAG not all words in student and model answers are equally important. Rather, pairs of related words which appear in student and model answers are more important than some other words. Hence, for evaluation of answers Document and Word vector based approaches are used.

Student answers often contain information beyond the key concepts instructors are looking for, though those extra pieces of text typically do not affect their scores unless they are contradictory or wrong.These shortcomings are taken care by Vecalign and Vecalign-asym.

Vector Representation in ASAG uses techniques like LSA(Latent semantic analysis),Paragraph vectors, Averaging word vectors,Word mover’s distance. This approach increases the precision of ASAG as it uses not only some keywords but uses a string or combination(pairs) of keywords while evaluation which overcomes the problem of dissimilarity between the representation student and provided model answers.

***C. An Iterative Transfer Learning Based Ensemble Technique for Automatic Short Answer Grading[3]:*** This work proposes an iterative technique which consists of two modules.

Module one consists of an ensemble which is a combination of two classifiers.

The first classifier uses TFIDF vectorisation on bag of words representations of student answers and then converts them to vectors with corresponding grades as class labels.

The second classifier is based on real valued features capturing similarity of student answers with respect to the model answer.

Module two consists of transfer learning based on canonical correlation analysis of a common feature representation to build the classifier ensemble for questions having no labelled data.

Projection vectors are used to transform the real valued features from the source question and target question to have maximum correlation. The source question’s labels are projected onto a subspace to train a model which is then used to predict labels of target question in this subspace.

This technique outperforms all the winning supervised entries on the SCIENTSBANK dataset from the “Student Response Analysis” task of SemEval 2013.

***D. Enhanced Bleu Methodfor Automatic Short Answer Grading[2]:*** This work proposes to assess an answer after calculating a score based on explicitly matching the student’s answer and the teacher's answer i.e. reference/model answer word by word. When the reference answers available are more than one the work will match them independently and the best scoring pair is taken as the final score.

Matching of unigrams are done based on the following modules:

A] Exact module: The module will match the surface level forms of unigrams.

B] Stemming module: The matching of two unigrams is done after stemming them down to their base form using Porter stemmer..

C] Heuristics Rule based module: The module matches unigrams based on their base form after applying the following heuristic rules:

Rule 1] WordNet synonym match: matches if the unigrams have the same parts of speech and are in the same synset of the WordNet.

Rule 2] Numeric value match: Matches the numeric value with the same written in text. (Eg. “2nd”is matches with “second”)

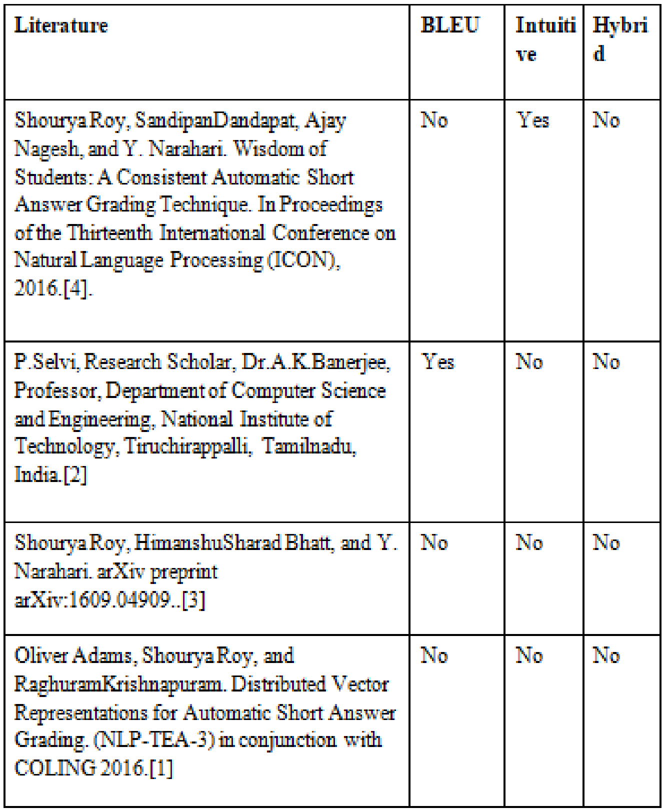
Rule 3] Acronym match: Matches the nodes with capitalised characters with the first characters of the corresponding multi word. (Eg. “NLP” is aligned with “Natural Language Processing")

Rule 4] Derivational form match: The Rule matches words which have the same root form or have a synonym with the same root form and which have similar semantic meaning, but which may belong to different syntactic categories.

Rule 5] Country adjectival form / demonym match: It matches from an explicit list of place names, adjectival forms, and demonyms.(Eg. “Chennai” and “Madras”)

### 2.1 Summary of Related Work

A literature review is an objective, critical summary of published research literature relevant to a topic under consideration for research. The summary is presented here in Table 1.



| Literature | BLEU | Intuitive | Hybrid |
|---|---|---|---|
| Shourya Roy, SandipanDandapat, Ajay Nagesh, and Y. Narahari. Wisdom of Students: A Consistent Automatic Short Answer Grading Technique. In Proceedings of the Thirteenth International Conference on Natural Language Processing (ICON), 2016.[4]. | No | Yes | No |
| P.Selvi, Research Scholar, Dr.A.K.Banerjee, Professor, Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, Tamilnadu, India.[2] | Yes | No | No |
| Shourya Roy, HimanshuSharad Bhatt, and Y. Narahari. arXiv preprint arXiv:1609.04909..[3] | No | No | No |
| Oliver Adams, Shourya Roy, and RaghuramKrishnapuram. Distributed Vector Representations for Automatic Short Answer Grading. (NLP-TEA-3) in conjunction with COLING 2016.[1] | No | No | No |

Table 1 Summary of literature survey

## III. PROPOSED WORK

The systems take input as student answers and model answers and then perform intuitive technique and the enhanced BLEU technique independent of each other. We then take a weighted combination of grades given by both the techniques to provide a final score or grade to the student answer. Scoring System: It would give the final score to the students by taking the weighted combination of the scores.

$$\text{Score} = IS*\alpha+ES*(1-\alpha) \qquad ....(3.1)$$

Where,
IS => Intuitive Score
Es => Enhanced BLEU Score
$\alpha$ => Weightage between (0,1)

### 3.1 System Architecture

The system architecture is given in Figure 1. Each block is described in this Section.

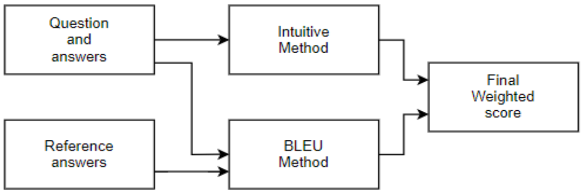


Fig. 1 Proposed system architecture

**Preprocessing Module:** This module performs all the preprocessing needed to be done on the input in order to perform NLP actions on them. This includes tokenization, word alignment, stop words removal etc.

**Enhanced BLEU Method:** This method assesses a text by computing a score based on explicit word-to-word match between the student's answer and teacher's answer (i.e. reference). If more than one reference is available, the matching similarity is scored against each reference independently and the best scoring pair is used to find the final score.

**Wisdom of Students- A Consistent Automatic Short Answer Grading Technique(Intuitive):** This technique makes two passes of the student answers. In the first pass it creates a skeleton of how the answers should be, based on the answers written by the students. Then it goes for a second pass in which it compares them with the created skeleton and scores them accordingly.

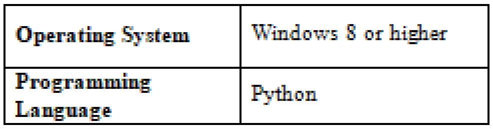
**Processing and Grading Module**: This module performs the main tasks of the algorithm that is heuristic rule matching and intuitive matching. It then computes the similarity between the model and the student answers and takes the weighted average between the two. After that scores are assigned based on the weighted average.

**Validation Module**: Validation module is required to compute the performance of the system. Correlation is calculated between the grades given by the teachers and grades given by the system to check real world performance of the system

## IV. REQUIREMENT ANALYSIS

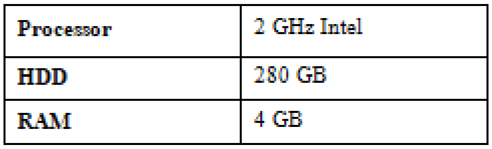
The implementation detail is given in this section.

### 4.1 Software



| | |
|---|---|
| **Operating System** | Windows 8 or higher |
| **Programming Language** | Python |

### 4.2 Hardware

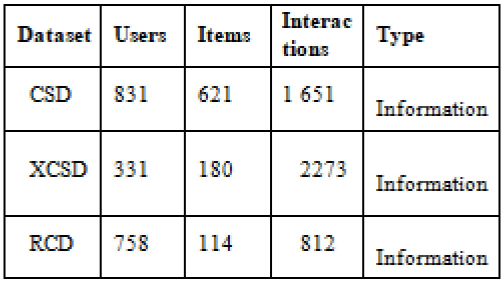


| | |
|---|---|
| **Processor** | 2 GHz Intel |
| **HDD** | 280 GB |
| **RAM** | 4 GB |

### 4.3 Dataset and Parameters

The AGSAS requires training of the model with an already given set of questions and answers to further grade the answers of unseen questions.The model is given the input of student answers for various questions, for example The CSD dataset consists of 21 questions with answers provided by a class of 31 students.

Table 3.3.1 Sample Dataset



| Dataset | Users | Items | Interactions | Type |
|---|---|---|---|---|
| CSD | 831 | 621 | 1 651 | Information |
| XCSD | 331 | 180 | 2273 | Information |
| RCD | 758 | 114 | 812 | Information |

An important aspect of AGSAS is to use appropriate evaluation measures for judging goodness of the automated techniques.

**Absolute error measure:** These measures judge the virtue of an automated technique based on the extent of element wise differences. Variations exist in terms of how differences are measured. Some of the popular examples are, Mean absolute error (MAE) and root mean square error (RMSE) which are shown to be superior measures than RMSE in assessing average model performance for classification tasks. MAE(u, v) is defined as the mean absolute differences between elements of u and v i.e.

$$\frac{1}{n}\sum_{i=1}^{n} \quad |u_i - v_i| \text{ ....(3.2)}$$

**Correlation coefficient:** A correlation coefficient is a number which is a quantification of some type of correlation and dependence. Pearson's r(u, v) is the most popular product-moment correlation coefficient between u and v where u¯ (v¯) denotes the mean of u (v) and σu (σv) denotes standard deviation of u (v).

$$r(u,v) = \frac{\sum_{i=1}^{n} \quad (u_i - \underline{u})(v_i - \underline{v})}{\sigma_u \sigma_v} \text{ ...(3.3)}$$

**Confusion matrix based measure:** A confusion matrix is generally used to evaluate supervised learning algorithms. Each column of the matrix represents the instances in actual class while each row represents the instances in an predicted class or vice versa. In the case of AGSAS, one can represent u and v along the rows and columns where principal diagonal elements indicate complete agreements, elements adjacent to the principal diagonal differ by 1 point and so on.

## REFERENCES

[1] Oliver Adams, Shourya Roy, and Raghuram Krishnapuram. Distributed Vector Representations for Automatic Short Answer Grading. In Proceedings of the Third Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA-3) in conjunction with COLING 2016.

[2] P.Selvi, Research Scholar, Dr.A.K.Banerjee, Professor, Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, Tamilnadu, India.

[3] Shourya Roy, HimanshuSharad Bhatt, and Y. Narahari. An Iterative Transfer Learning Based Ensemble Technique for Automatic Short Answer Grading. arXiv preprint arXiv:1609.04909.

[4] Shourya Roy, SandipanDandapat, Ajay Nagesh, and Y. Narahari. Wisdom of Students: A Consistent Automatic Short Answer Grading Technique. In Proceedings of the Thirteenth International Conference on Natural Language Processing (ICON), 2016.

[5] Shourya Roy, Y. Narahari, and Om D. Deshmukh. A Perspective on Computer Assisted Assessment Techniques for Short Free-text answers. In Proceedings of the International Conference on Computer Assisted Assessment (CAA), pages 96-109. Springer, 2015.