International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Intelligent Tutoring System for Evaluating Student Performance in Descriptive Answers Using Natural Language Processing

Shweta Patil[1], Sonal Patil[2]

[1]M.E. (CSE) Student, Department of CSE, G.H.Raisoni Institute of Engineering and Management Jalgaon, Maharashtra, India

[2]Assistant Professor, Department of Computer Engineering, G.H.Raisoni Institute of Engineering and Management Jalgaon, Maharashtra, India

Abstract: ***Computer Assisted Assessment of free-text answers has established a great deal of work during the last years due to the need of evaluating the deep understanding of the lessons' concepts that, according to most educators and researchers, cannot be done by simple MCQ testing. In this paper we have reviewed the techniques underpinned this system, the description of currently available systems for marking short free text response and finally proposed a system that would evaluate the descriptive type answers using Natural Language Processing and lastly compared the results obtain by human graders and proposed system. We have also compared the results of proposed system with existing systems.***

Keywords: Computer Assisted Assessment, Short free text response, Descriptive type answer, Natural Language Processing.

1. Introduction

"Computer Assisted Assessment (CAA) is a common term for the use of computers in the assessment of student learning [1]". The idea of using computers to assist learning process has surprisingly changed the field of learning system. The study in the field of CAA started nearly in 70's. The CAA systems developed so far are capable of evaluating only essay and short text answers such as multiple choice questions, short answer, selection/association, hot spot and visual identification. Most researchers in this field agree on the notion that some aspects of complex achievement are complicated to measure using objective type questions. Learning outcomes implying the ability to recall, organize and integrate ideas, the ability to express oneself in writing and the ability to supply merely than identify interpretation and application of data, require less structuring of response than that imposed by objective test items [5]. Due to these students will be evaluated at higher level of Bloom's (1956) Taxonomy (namely evaluation and synthesis) that the essay question or descriptive questions serves it's most useful purpose.

Many researchers claim that the essay's evaluated by assessment tools and by human graders leads to great variation in score awarded to students. Also many evaluations are performed considering specific concepts. If that particular concepts are present then only award grades otherwise answer is marked as incorrect. So to overcome this problem new system is proposed.

Purpose of this paper is to present the new system that can evaluate student's performance at higher level of Bloom's taxonomy by considering the assessment of descriptive type questions. The system can perform grading as well as provide feedback for student's to make improvement in their performance. This paper also discusses various techniques underpinned by computer assisted assessment system as well as current approaches of CAA and utilizes it as a framework for designing our new framework.

The techniques for automatic marking of free-text responses are basically categories into three main kinds, Statistical, Information Extraction and Full Natural Language Processing [2].

A. Statistical Technique

It is only based on keyword matching, hence considered as poor method. It cannot tackle the problems such as synonyms in student answers, nor does it takes into account the order of words, nor can it deal with lexical variability.

B. Information Extraction (IE) Technique:

Information Extraction consists in getting structured information from free text. IE may be used to extract dependencies between concepts. Firstly, the text is broken into concepts and their relationships. Then, the dependencies found are compared against the human experts to give the student's score.

C. Full Natural language processing (NLP):

It involves parsing of text and finds the semantic meaning of student answer and finally compares it with instructors answer and assigns the final scores.

2. Related Work

Jana and John developed C-rater [3] to score student's answers automatically for evidences of what a student knows about the concepts already described by C-rater. It is underpinned by NLP and Knowledge Representation (KR) techniques. In this system model answers are generated with the help of concepts already given and later student's answers are processed by NLP technique. Later on only concepts detection is done and finally scores are assigned. But there are some disadvantages of this system

as No distinct concepts specified, incorrect spelling mistakes unexpected similar lexicons and many more. J.Burstein, K.Kukich developed E-rater [8] in 1998 to automatically analyze essay features based on writing characteristics specified by six different score points in the scoring guide used by human graders for manual scoring. E-rater features include rhetorical structure, Syntactic structure and topical analysis. Then Raheel, Christopher and Rosheena created IndusMarker [4] an automated short answer marking system for Object Oriented Programming (OOP) course. The system was used by instructors to assist the overall performance of student and to provide feedback to the students about their performance. It exploits the structure matching i.e. matching the pre specified structure with the contents of student response text. Automated Essay Grading (AEG) system was developed by Siddhartha and Sameen in the year 2010 [5]. The aim of the system is to overcome the problems of influence of local language in English essays while correcting and by giving correct feedback to writers. AEG is based on NLP and some of the Machine Learning (ML) techniques. Auto-assessor was developed in year 20111 by Laurie and Maiga [6] with an aim to automatically score student short answers based on the semantic meaning of those answers.Auto-assessors is underpinned by NLP technique. This system consists of component based architecture. The components are created in order to reduce the sentences to their canonical form which are used in preprocessing of both supplied correct answers as well as student response. Later evaluation of student response with the correct answer takes place where each word from correct answer in canonical form is compared with the word from student response which is in canonical form and finally scores are awarded for student response. Ade-Ibijola, Wakama and Amadi developed Automated Essay Scoring (AES) an Expert System (ES) for scoring free text answers[7]. AES is based on Information Extraction (IE). This ES is composed of three primary modules as: Knowledge Base, Inference Engine and Working Memory. Inference engine uses shallow NLP technique to promote the pattern matching from the inference rules to the data contained in knowledge base. The NLP module contains: a Lexical Analyzer, a Filter and a Synonyms Handler module. The correctness evaluation is performed by fuzzy model which generate the scores for student answer with the help of two parameters: the percentage match and the mark assigned. Later P.Selvi and A.K.Bnerjee developed automatic short answer grading system [9] which was enhanced by BLUE method. ASAGS basically consist of preprocessing, mapping, feedback and validation module. In preprocessing student answer is converted into XML format by breaking text into tokens and identifying sentence boundaries. In Mapping module unigrams from student answer are mapped with model answer based on exact, stemmed and heuristic rules. In feedback module numerical scores are awarded to student answer and finally in validation module human scores and system scores are compared and the correlation between human and system scores are computed.

3. Problem Definition

There are a number of commercial assessment tools available on the market today; however these tools support objective type question such as multiple choice Questions or short one-line free text responses. This will assess student's depth of knowledge only at lower level of Blooms taxonomy of educational objectives. They fail to assess student's performance at higher level of taxonomy of educational objective. Also these systems fail to check spelling & grammatical mistakes made by students. As well as they were unable to check the correct word order. Even the answers with wrong word order were awarded assigned scores by mere presence of words in student response. So to overcome the encountered problems the system is going to be developed that evaluated students descriptive answers by considering the collective meaning of multiple sentences. Also system will mark spelling mistakes made and finally scores will be assigned to student answer. The proposed system will try to provide feedback to the students by allotting numerical scores so to help them to improve their performance in academics. Finally the proposed system is developed to overcome the limitations of the existing systems mention above it will make the evaluation of descriptive type answers with accurate and timely results.

4. Proposed System

The proposed system will implement CAA for descriptive type answer. The existing system checks single line text response without considering word order. So the proposed system will try to avoid this problem by considering collective meaning of multiple sentences. The primary focus of this newly proposed system is to determine the semantic meaning of student answer with a consideration that student responses to question in number of ways. The system basically focuses on multiple sentences response.

The basic architecture of proposed system is depicted in fig [1] below. It is basically composed of following components:

a) Student Module:
It consists of question editor where question will be displayed and response editor to enter student response.

b) Tutor Module:
In this module question as well as correct response to respective question is entered by tutor. Tutor will also identify and enter the keywords from correct answer with their respective weights.

c) Processing Module:
Both answers i.e. student response and correct answer will be processed by initially dividing them into token i.e. words. Later on noun phrase and verb grouping will be assigned to each and every word with the help of Part-Of-Speech (POS) tagger. This task is accomplished by NLP technique.

d) Answer comparison and Grade Assignment Module:
Following text processing module, that actual evaluation of student response with correct answer takes place. Each and every word of student response is compared with

correct answer. If exact match is found in word as well as POS tag and word position in sentence the scores are assigned. After score assignment Final scores are calculated by making summation of assigned scores of all words.

e) Projection of Final Scores:
Final calculated scores assigned to student response are given in report.

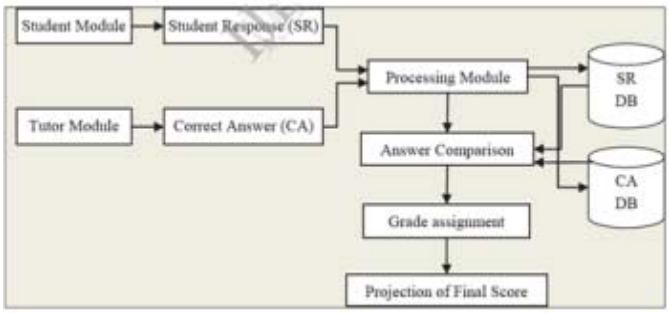

Figure 1: Architecture of Proposed System

Now steps for evaluating the descriptive type answer of the proposed system are:

- *Step 1:* Start.
- *Step 2:* Form the correct answer and store all the words present in it in a master table.
- *Step 3:* Identify and tag the key words and key verbs. Tag other words as the non-key terms. Put weights to all the words according to their importance in the answer. Calculate the weights for actual correct answer by adding all the assigned scores of all words present in it and store it.
- *Step 4:* Collect and insert all the probable synonyms and antonyms of the words of the correct answer to a synonym table and antonym table.
- *Step 5:* Assign a weight to each synonym considering the change of meaning of the sentence due to its presence.
- *Step 6:* Set student's score for the answer to 0. Input student's answer. Split the input into words and store them in a temporary table.
- *Step 7:* Check if the key words or key terms are present in the temporary table. If key words or key terms found put the weight for that word. score = score + new weight. Go to *Step 11*.
- *Step 8:* If key word or key verb not found then check if any synonym word of the key terms are found. If synonyms for key terms found put the weight for that word. Score = score+ new weight. Go to *Step 11*.
- *Step 9:* If synonyms for key terms not found then stop further checking and consider the answer as ERROR.
- *Step 10:* Check the position vectors of the nouns and verbs combination in the input answer and compare it to that of the correct answers to verify the dependencies of the nouns and the verbs in the answer.
- *Step 11:* Put the weights of the non-key terms accordingly. Put weight 0 for any unknown word found. Now calculate the net weight.
- *Step 12:* If the net score is negative the result is an FAIL. If the net weight is positive and in the range of original scores then the result is PASS.
- *Step 13:* End.

5. Experiment Results

5.1 Evaluation

The testing was performed in a real world classroom scenario by administering a test of five (5) subjects to a student- making a totality of 30 test samples. The student responses were then assessed by the newly developed assessment system and also by a subject expert. The scores obtained from both assessments are presented in Table 1 ('a' to 'e') below

Table 5.1: Subject Expert (SE) scores and proposed Assessment System (AS) scores

SE	3	5	8	10	7
AS	5	4	6	10	6

(a) System Programming

SE	5	2	4	2	2
AS	4	2	4	3	2

(b) Linux

SE	6	2	3	3	0
AS	5	2	2	3	2

(c) Java

SE	5	3	3	4	5
AS	5	2	2	4	6

(d) C-programming

SE	4	3	10	5	2
AS	3	3	8	3	2

(e) Artificial Intelligence

SE	4	3	5	3	3
AS	3	3	5	2	4

(f) Computer Network.

Several metrics have been proposed, adopted and/or used for evaluation of descriptive type answers. [1]Some notable metrics are: Measure of Exact Agreement, Adjacent Agreement, the Pearson Correlation, Spearman or nonparametric Correlation, Mean and Standard Deviations, Kappa Measure and F-Score. However, we have used Pearson Correlation Coefficient for the proposed system.

5.2 Pearson Correlation

It measures the standard correlation, that is, how much the teachers scores or true scores (X) are related with the systems scores (Y). It is calculated by applying Eq. (1) below. It is suitable whenever answers are evaluated with numerical scores. Sometimes the true scores are the results of the average consensus of several teachers.

$$\text{Correlation}(X,Y) = \frac{\text{Covariance}(X,Y)}{\text{Standard Deviation}(X) * \text{Standard Deviation}(Y)} \quad \text{Eq (1)}$$

The Correlation Coefficient obtained from the computation is 0.8. A graphical description of this correlation is presented in fig 8. The overlapping lines on the graph indicate the number of times the teachers score was exactly the same as the systems score.

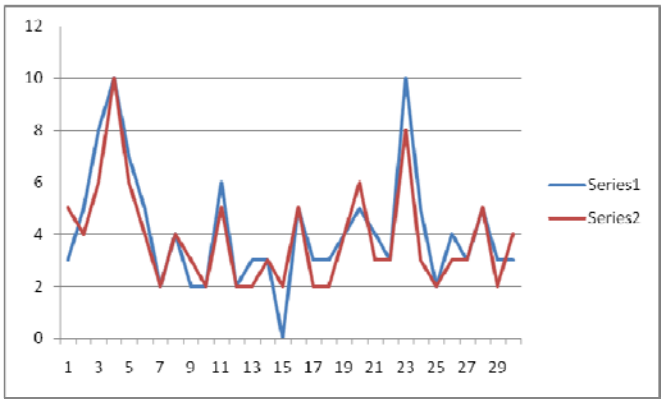

Fig 5.14 Correlation between Subject Expert scores and Assessment System scores Using 30 Test Samples

And later on the comparative study was done with the existing 4 different automated assessment systems. Table 2 gives the overview of the techniques, evaluations and languages of the reviewed Automatic assessment system.

Table 5.2: Overview of the techniques, evaluation and languages of reviewed AES system

System	Techniques	Evaluation	Languages
Proposed System	NLP	Corr:0.78	English
C-rater	NLP	Arg:0.83	English
E-rater	NLP and VSM	Agr:0.97	English
Larkey's system	TCT	EAgr:0.55	English
ES4ES	IE	Corr:0.71	English

6. Conclusion and Future Scope

CAA has been an interesting research area since70's. CAA helps in evaluation of student performance accurately and without wastage of time. Most of the CAA developed provides promising results compared with results provided by human graders. This work presents an assembled approach with a number of essential features of an ITS including a metric embedded to it. The metric helps improving the evaluation of descriptive type test sessions. The evaluation is done through test sessions on various subjects. Feedbacks given are explicit and suggestive. The metric of evaluating the descriptive answers works best for simple multiple sentences. The provision to enter new words in the databases makes the databases flexible. As it is not completely dependent on spell checking and grammar checking the result may not always be satisfactory. The word dictionary and databases used in the metric is upgraded for more accurate results. The results produced will increase the performance by grading mostly accurate results. The results produce will be near to the grades produce by the human grader.

The system is even tested and the results projected by the system are compared with the grades allocated by human grader. The system is also compared with the systems developed so far in terms of correlation. In future plug-ins can be used to develop well-formed spell checker and grammar checker to obtain better results. With an improved grammar checker the metric can also support compound and complex answers.

References

[1] Perez D. 2007, "Adaptive Computer Assissted Assessment of Free Text Students answers: an approach to automatically generate students conceptual models." PhD Thesis, computer science Department Universidad Autonoma de Madrid, 2007.

[2] Mitchell T, Russell Broomhead P, and Aldridge N. 2002, "Towards robust computerised marking of free text responses." In proceeding of the 6th computer Assissted Assessment Conference, 2002.

[3] Jana Z. Sukkarieh, and John Blackmore," c-rater:Automatic Content Scoring for Short constructed Responses." In proceeding of the 22nd International FLAIRS Conference,2009.

[4] Raheel Siddiqi, Christopher J. Harrison and Rosheena Siddiqi,"Improving Teching and Learning through Automated Short- Answer Marking." IEEE Transactions on learning technologies, vol.3, No.3, July-September 2010.

[5] Siddhartha Ghosh and Dr. Sammen S Fatima,"Design of an Automated Eassy Grading (AEG) system in Indian Context." Internation Journal of Computer Application (0975-8887), vol.1, No.11, 2010. Laurie Cutrone, Maiga Chang and Kinshuk,

[6] "Auto-Assessor: Computerized Assessment System for Marking Student's Short-Answers Automatically", IEEE International Conference on Technology for Education,2011.

[7] Ade-Ibijola, Abejide Olu, Wakama, Ibiba, Amadi, Juliet Chioma, " An Expert System for Automated Essay Scoring (AES) in Computing using Shallow NLP Techniques for Inferencing", International Journal of Computer Applications (0975 - 8887) Volume 51-No.10, August 2012.

[8] Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lut Martin Chodorow, Lisa Braden-Harder$:l:, and Mary Dee Harris:H:, *"Automated Scoring Using A Hybrid Feature Identification Technique"*, Proceeding Annual meeting of the association of Computational Linguistic (ACL 1998), August 10-14 1998, pp 206-210.

[9] P.Selvi, DR.A.K.Bnerjee,*"Automatic Short-Answer Grading System (ASAGS)"*, InternationalJRI Computer Science and Networking, Vol.2, Issue 1, August 2010, pp-19-23.

Author Profile

Shweta Patil has received her degree in Computer Engineering from SSBT's College of Engineering and Technology, Bambhori, Jalgaon 2011. She is currently pursuing her Masters of Engineering in Computer Science and Engineering from G.H.Raisoni Institute of Engineering and Management, Jalgaon under North Maharashtra University, Jalgaon.

Sonal Patil received B.E. degree in Computer Engineering from SSBT's College of Engineering and Technology, Bambhori, Jalgaon from North Maharashtra University in 2008 and M.Tech in CSE from TIT, Bhopal from Rajiv Gandhi Prodyogiki Vishvadalaya in 2012. She is currently working as a Assistant. Professor in G.H.Raisoni Institute of Engineering and Management, Jalgaon. She has published 7 article in National and International Journal and 17 papers in National and International Conferences Member of ISTE.