# HYBRID APPROACH FOR AUTOMATIC SHORT ANSWER MARKING

**Shaha T. Alotaibi**
Information Systems Department, Princess Nora Bint Abdul Rahman University, Riyadh, Saudi Arabia, shturki@hotmail.com

**Abdulrahman A. Mirza**
Information Systems Department, King Saud University, Riyadh, Saudi Arabia, amirza@ksu.edu.sa

## ABSTRACT

*E-assessment is a key element in any e-learning system, needed to evaluate the learning process. It can be successfully and easily carried out on Multiple Choice Questions (MCQs). However, e-assessment of essay questions is much harder than that of MCQs. Consequently, it is a growing area of research. This paper presents an approach to assess short answer questions automatically through integrating the Information Extraction (IE) technique and Decision Tree Learning (DTL), a Machine learning (ML) technique. The IE technique employs Natural Language Processing (NLP) tools such as parsers and lexicon. Additionally, the Machine Learning techniques automate the free-text marking using the classification rules that are extracted from DTL and used to assess the student's answer.*

Index Terms – E-assessment, E-learning, NLP, IE, DTL.

## 1. INTRODUCTION

Technology today presents many novel opportunities for innovation in the student's answers assessment through new assessment tasks and potentially powerful scoring and reporting mechanisms. E-assessment has grown exponentially in the last few years, due to an increasing number of online students and the orientation towards learning environments based on Information and Communication Technology (ICT). Electronic assessment is defined as: "the use of computer to set, deliver and often mark tests of student skill, understanding and knowledge of subject" (Intelligent Assessment, 2011). Most learning management systems provide assessment templates for Multiple-Choice Questions (MCQs). However, essay questions marking is limited in these systems. The students in many cases respond to the MCQs through guessing. Furthermore, MCQs that assess the higher-order thinking skills are hard to build, and time consuming. Hence, essay questions can be a good alternative to evaluate such skills. In fact, exceeding the limitations of technology in automatic grading of these types of questions benefit the learning process (Callear, et. al, 2001; Govindasamy, 2002).

The essay questions can be divided into two categories: long answer and short answer questions. The long answer is a free text where the students talk about a certain subject. This type is graded by evaluating the writing style and the contents. The short answer questions typically request that students write one or two sentences (Sukkarieh, et. al, 2003).This type of questions is graded by evaluating the content where the style is not necessary (Williams &

Dreher, 2004) We focus in this paper on automatic short answer grading which is needed in scientific disciplines and still a challenge in e-assessment.

We propose a method that integrates Information Extraction (IE), and Decision Tree Learning (DTL), a ML technique. IE enables the use of NLP tools (parsers, lexicon, etc.) (Mitchell, et. al, 2002). However, ML techniques can automate free text marking without having to develop systems that totally understand the student response (Pulman & Sukkarieh, 2005).

The rest of this paper is organized as follows: Section 2 presents the related work. Section 3 describes the proposed approach. Section 4 presents the marking algorithm. In section 5, the training and test materials are applied to our algorithm. Section 6 discusses some issues related to the proposed approach. Finally, a conclusion of this paper in Section 7.

## 2. RELATED WORK

Several approaches have been proposed to automate assessment. They can be grouped into four main categories: NLP, IE, classification, and integrated methods.

NLP is an application of computational methods that are used to analyze characteristics of electronic files of text or speech. Methods used are either statistical or linguistic based analyses of language features. It employs tools such as syntactic parsers, to analyze the syntactic form of a text; discourse parsers, to analyze the discourse structure of a text; lexical similarity measures, to analyze word use of a text (Burstein, et. al, 2002). As an example, C-rater is a NLP system which is developed by ETS technologies. This application evaluates the understanding of content materials by mapping the answers onto a model and then determining the correctness or incorrectness of the student's answer (Leacock & Chodorow, 2003).

IE is a type of information retrieval technique that aims to automatically extract structured information from unstructured and/or semi-structured machine-readable documents. AutoMark is an automatic assessor developed by the Intelligent Assessment Technologies in the UK. It uses IE techniques to provide automatic marking. AutoMark searches for specific content in the student's response. The content determined in the form of a set of templates. Each template represents a correct or incorrect answer. Student answers are first parsed, and then intelligently matched against the template, and a mark for each answer is computed ((Mitchell, et. al, 2002; Mitchell, et. al, 2003).

Furthermore, various studies have examined ML approaches. ML is a branch of artificial intelligence that is concerned with the development of algorithms allowing the machine to learn via inductive inference based on observing data that represent incomplete information about statistical phenomenon (Mitchell, 1997). Classification techniques are an important task in ML by which machines learn to automatically recognize complex patterns. For example, the use of DTL and Naive Bayesian learning (Nbayes) in marking problems (Pulman & Sukkarieh, 2005).

BETSY which is a program developed at the University of Maryland classifies text based on trained material. This system determines the most appropriate classification using a large set of features. Each text is viewed as a particular case of all the calibrated features. The probability of each score for a certain text is calculated as the product of the probabilities of

the features that are contained in the text. Then, the conditional probability of existence of each feature is predicated by the proportion of texts within each category that includes the feature (Valenti, et. al, 2003; Rudner & Liang, 2002).

Finally, the integrated techniques aim to combine the strengths of multiple approaches while avoiding some of the weaknesses. CarmelTC, is a hybrid text classification approach for analyzing essay answers of qualitative physics questions. It classifies pieces of text based on the extracted features from a syntactic analysis, as well as on a Nbayes classification of the same text (Rose, et. al, 2003)

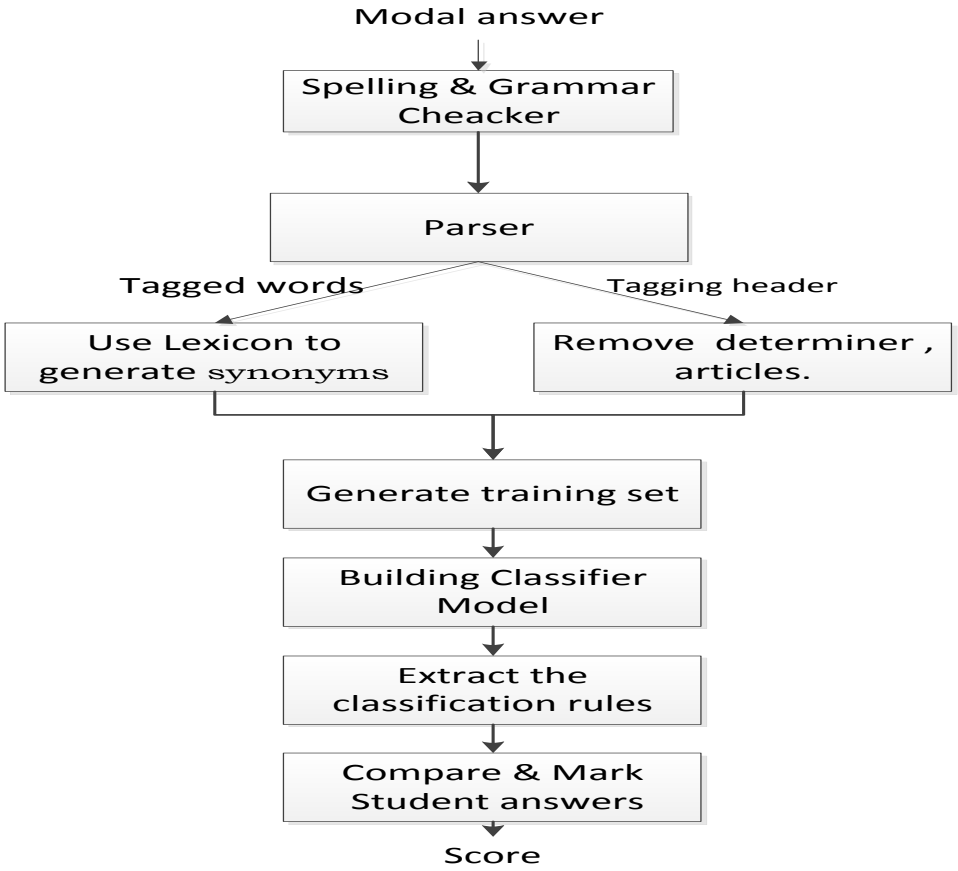
## 3. PROPOSED APPROACH

As mention above, several techniques exist for dealing with e-assessment. We propose a method that integrates Information Extraction (IE), and Decision Tree Learning (DTL) technique. IE enables the use of NLP tools (parsers, lexicon, etc.) (Mitchell, et. al, 2002). However, ML techniques can automate free text marking without having to develop systems that totally understand the student response (Pulman & Sukkarieh, 2005).

The proposed approach is mainly composed of four steps: parsing, training set building, learning, and classifying steps. First, the examiner writes the model answer for the question. Multiple templates of model answers may be written. The model answers are preprocessed to normalize the input in terms of spelling and grammar. The result is fed into the parser such as the Stanford Parser. The tagging headers are used as attributes, for example NN, VBZ, NNS, IN, which are, respectively, tags for a singular noun, singular present verb, plural noun, and preposition. The tagged words from a model answer are used as an instance for the attributes. The student's answer also follows the same process as the model answers.

Second, different synonyms for each tagged word are generated to build different training sets. Additional words are added to each group of synonyms. For example, the value "X" can be used to build incorrect training sets. Each set containing "X" in any attribute is classified as "Incorrect", otherwise; it is classified as "Correct". If there is more than one template of the model answer, then the same process is applied to each template. Third, One decision tree is built for each template of model answers. All decision trees will be used to extract classification rules. Fourth, the tagged words from the student's response are compared with the set of rules. The counter is used to count number of matched words in the student's response. If the counter equals the token (number of tagged words) of the model answer, then the student is given the full-mark that determined by the examiner, otherwise; the student will get a mark equal to the counter/token*full-mark. This equation will give a percentage of full-mark depends on the correctness of the student's answer. If the answer contains more than one sentence, then the procedure is applied to each sentence and the mark is divided between the sentences. Fig.1 depicts the system architecture of this approach.

**Figure 1: System Architecture of the proposed approach**




## 4. MARKING ALGORITHM

For each question, the training set is generated and then, the set of rules is extracted. The words set of rules stored in matrix **R** [number of rules, number of tagged words]. The letters **i** and **k** are both integers that used as indices for the matrix ***R***. **S** is the set of words and **W** is one word in the student's response. The **In_counter** and **C-counter** are respectively, number of incorrect words and number of correct words in the student's response.

```
Read Rules and load words in Matrix R
Given Student response and store it in S
'Classify Student's answer
C_counter =0
In_Counter =0
Repeat until no more rules
  Repeat until no more Words in the current rule
   If  W of S unmatched R[i,k]
         if R[i,k] = 'X'
           Match W with synonyms words for this attribute
           If W Unmatched synonyms words
            W replaced by 'X' in S
  Take next word in S
Go to next rule
```

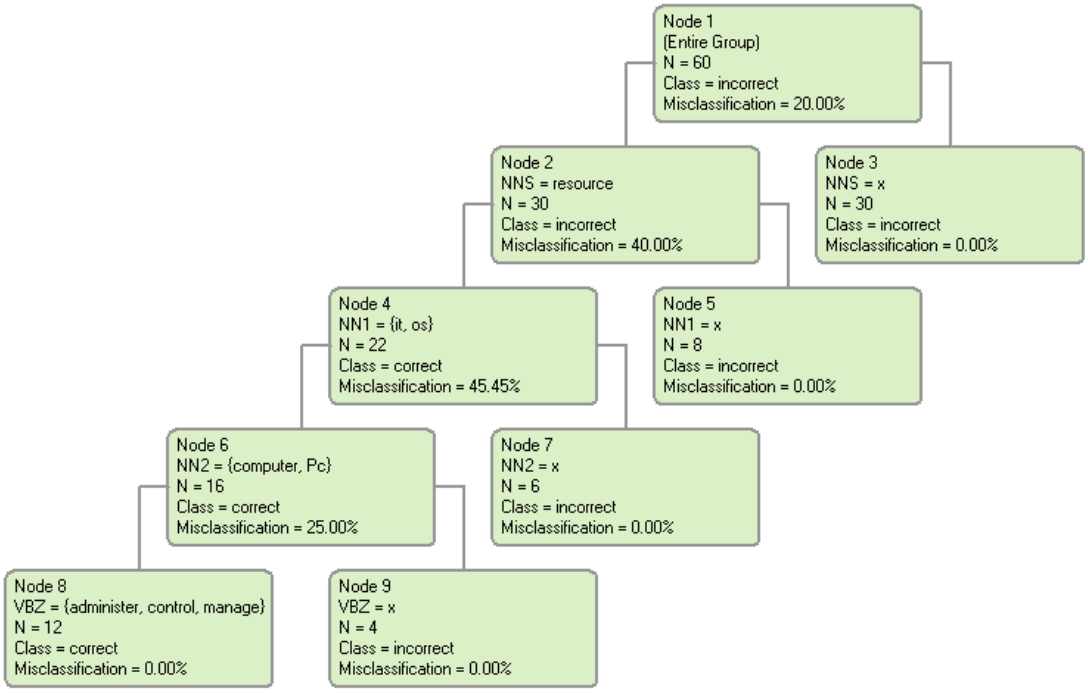
*'Marking*
*Count number of 'X' in S and put value in In_counter*
*C_counter = Token – In_counter*
*If C_Counter = Token Then*
*Score = full-mark*
*Else*
*Score = C_Counter / Token * full-mark*
*'Feed back*
*Given S after change wrong words to 'X'*
*Find X in S*
*Attach the appropriate feedback*
*Display the result with feedback message*

## 5. TRAINING AND TEST MATERIALS

The following sample question is a simple question taken from a computer introduction course; where the questions of computer field are limited to few sentences and dealt with explicit concept rather than subjective opinions that are needed to interpret.
Question: What is the main function of OS?
Model answer: The OS manages the computer resources.

1. After spelling and grammar are checked, the next step is using of Stanford Parser that produces the following output:
   ***Tagging words:*** The/DT**,** OS/NN**,** manages/VBZ**,** the/DT**,** computer/NN**,** resources/NNS**.**
   ***Tokens***: 6
2. Discard the determiners, articles and other words which have very low discrimination power. For any word removed, one is subtracted from the token.
   ***Tagging*:** OS/NN, manages/VBZ, computer/NN, resources/NNS
   ***Tokens***: 4
3. Return any verb to its head and any plural noun to the singular noun.
   ***Tagging*:** OS/NN, manage/VBZ, computer/NN, resource/NNS
   ***Tokens***: 4
   If there is more than one element, which has the same tag; numbers are used to differentiate between them. The noun tag appears two times in the last result. Then the noun tags become NN1 and NN2.
4. Generate the synonyms for each tagged word. In our example, the "OS" word can be replaced by "It". The verb "manage" can be replaced by "control" or "administer". The "computer" word can be replaced by "PC". We used the value "X" to represent the wrong words which is appeared in the student's response. Fig. 2 depicts the resultant decision tree for our example that is generated using the DTREG software.

**Figure 2: The resultant Decision tree**




5. Extract the classification rules from the resultant decision tree. The words of rules are stored in matrix and used to classify the student's response. The examiner can write multiple templates for this question to satisfy different writing styles. If there are many templates for the same question, a separate decision tree is built for each template. Next, the classification rules are extracted from all decision trees. The passive form can be written for the same answer of above question as follows:
   The computer resources are managed by the OS.
   ***Tagging***
   computer/NN, resources/NNS, are/VBP, managed/VBN, by/IN, OS/NN
   ***Tokens***: 6
6. Compare the student answer with the set of words of all rules that resulted from all decision trees. The student response classified immediately to "Correct" or "Incorrect". If the student answer classified as correct, the student takes the full mark, otherwise; the system will assign score equal to number of matched words/token * full mark. Additionally, the feedback is attached with the student's score depending on the wrong words in their response.

   The following responses will match the above templates:
   *The OS manages the computer resources.*
   *The OS controls the computer resources.*
   *The OS administers the computer resources.*
   *The OS manages the PC resources.*
   *The OS controls the PC resources.*
   *The OS administers the PC resources.*
   *It manages the computer resources.*
   *It controls the computer resources.*
   *It administers the computer resources.*

*It manages the PC resources.*
*It controls the PC resources.*
*It administers the PC resources.*
*The computer resources are managed by the OS.*
*The computer resources are controlled by the OS.*
*The computer resources are administered by the OS.*
*The PC resources are managed by the OS.*
*The PC resources are controlled by the OS.*
*The PC resources are administered by the OS.*

## 6. DISCUSSION

We suppose that the approach which obtains a prediction result based on the integration of two techniques would give better results than either of these techniques alone. Our approach aims to take advantage of both IE, as well as classification techniques. The required resources of this approach are easy to assemble. It includes a parser and a dictionary of synonyms joined with a decision tree builder. All these resources need to be integrated in a stand-alone system. Additionally, this method can be applied to many sentences by processing an individual sentence and comparing the student's response with the resultant rules. Furthermore, the marking scheme assigns a mark depending on the percentage of correctness in the student's response not just classifying the resultant mark as correct or incorrect.

In addition, determining the wrong words in the student's answer will help to provide a meaningful feedback, which is prepared in advance by the examiner. The use of various templates for each question increases the possibility to meet different writing styles that may be used by students, as well as using a spelling and grammar checker in the preprocessing stage, will help to avoid parsing errors. Returning verbs and nouns to their base and discarding the determiners, articles, and other words which have extremely less discrimination power helps to build simple decision trees and reduce the variation between the model answer and student's response.

On another hand, some limitation may accompany this approach in the case of bad-formed structures resulting from the writing styles of students. In addition, the use of shifting role, such as shifting of subject and object when using the passive voice. The last problem can be avoided by using additional templates of model answers for passive form when writing templates. In general, this approach targets the scientific disciplines where the sentences are explicit, such that there is no need for inference to understand the concepts.

## 7. CONCLUSION

This research proposes the combination of different techniques to produce an approach for automatic short answer marking. In particular we have presented an integrated method using Information Extraction (IE) and Machine Learning (ML) techniques. The IE enables the use of NLP tools such as parsers and lexicon. In addition, the ML techniques automate the free-text marking using DTL. The classification rules are extracted from DTL and used to assess the student's answer. An algorithm of the proposed approach is presented, and the training set is generated for sample question with a set of examples. Generating multiple templates for each question, as well as the preprocessing step of the input sentences for both the model

answer and the student response, increase the possibility to meet different writing styles that may be used by students.

## REFERENCES


Burstein, J., Leacock, C., & Swartz, R. (2001). Automated evaluation of essays and short answers. In Proceedings of the 6th International Computer Assisted Assessment Conference, Loughborough, UK.

Callear, D., Jerrams-Smith, J. and Soh, V. (2001), CAA of Short Non-MCQ Answers. 5th International Computer Assisted Assessment Conference, Loughborough, UK.

Govindasamy, T. (2002), Successful Implementation of E-Learning: Pedagogical Consideration. Elsevier Science Inc. PII: S1096-7516(01)00071-9.

Intelligent Assessment (2011), www.intelligentassessment.com.

Leacock, C., and, Chodorow, M. (2003), C-rater: Automated Scoring of Short-Answer Question, Computers and Humanities, 37 (4), 389-405.

Mitchell, T., Russell, T., Broomhead P., and Aldridge, N. (2002), Towards Robust Computerised Marking of Free-Text Responses. Proceedings of the 6th International Computer Assisted Assessment Conference, Loughborough, pp. 233-249.

Mitchell, T., Russell, T., Broomhead P., and Aldridge, N. (2003), Computerized marking of short-answer free-text responses. 29th Annual Conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

Mitchell, T. (1997), Machine Learning. McGraw Hill Publishing, New York, USA.

Pulman, S. G., and, Sukkarieh, J. Z. (2005), Automatic Short Answer Marking. in Association for Computational Linguistics: Proceedings of 2nd Workshop on Building Educational Applications Using NLP, Ann Arbor, Michigan, pp. 9-16.

Rose, C. P., Roque, A., Bhembe, D. and VanLehn, K. (2003), A hybrid text classification approach for analysis of student essays, in Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing, PA, USA , (2), pp. 68-75 .

Rudner, L.M. and Liang, T. (2002), Automated Essay Scoring Using Bayes' Theorem. The Journal of Technology, Learning and Assessment, 1(2), pp. 3-21.

Sukkarieh, J. Z., Pulman, S. G. and Raikes, N. (2003), Auto-marking: using computational linguistics to score short, free-text responses. 29th Annual Conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

Valenti, S., Neri, F., and Cucchiarelli, A. (2003), An overview of current research on automated essay grading. Journal of Information Technology Education, (2), pp. 319–330.

Williams, R. and Dreher, H. (2004), Automatically grading essays with Markit©. In Proceedings of Informing Science, Rockhampton, Australia.