© 2018 IJSRST | Volume 4 | Issue 2 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Handwritten Short Answer Evaluation System (HSAES)

Sijimol P J, Surekha Mariam Varghese

Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India

ABSTRACT

This paper presents the design and implementation of an automated evaluation system for short answers. Handwritten Short Answer Evaluation System (HSAES) is an automated short answer evaluation system that is capable of identifying the texts in answer papers and evaluates marks for each short answer based on previous knowledge acquired by the model. Although many essay evaluation systems are available, short answer grading is still a big problem. In the proposed system, Optical Character Recognition tools are used to extract the handwritten texts. Natural Language Processing is used to extract the keywords from human evaluated sample dataset of handwritten answer papers and answer key. The proposed model evaluates scores based on cosine sentence similarity measures. Each sentence in the evaluated answer paper carries their corresponding mark. The developed model can be used to evaluate the marks of the unscored short answers.

Keywords : OCR, Short Answer, NLP, Machine Learning, Scoring

I. INTRODUCTION

The writing quality or ability of a student can be improved by receiving feedback from the teacher. In recent years, automated essay scoring has become a hot issue in the research of natural language processing with the growing need of essay scoring in English writing skill. In many cases, the essay scoring task costs huge human resources but with less efficiency and the score given by human evaluator is mostly determined by his knowledge, emotion and energy. There may be a huge deviation between the scores evaluated by different raters. Even the same rater probably gives different scores for the same essay at different times. Thus, the correctness of essay scoring cannot be guaranteed. So answer paper evaluation and scoring can be tedious and time consuming process for many of evaluators. For many teachers to finish the essay scoring of all student essays in a short interval time is the major challenging task. Thus, students cannot feedback on their answers in time. Thus in order to solve these issues the researchers have been proposed automated essay scoring techniques. They analyses a piece of text based on its semantics, context and spelling. Although many essay evaluation systems are available, short answer grading is still a big problem.

In this paper we propose a model to automatically evaluate the short answers in the handwritten answer scripts. Optical Character Recognition tools are used to extract the sentences in the short answers and are converted to text files. Among them randomly chosen text files are used for evaluating the scores by the human evaluators and the evaluated short answer text files are used during training to develop a model. A high weightage given answer key text file is also used during training. Each sentence in the text has its corresponding mark. The semantic similarity between the sentences can be calculated by a cosine similarity based approach. During testing each unscored short answer text files can be used as input to the developed model.

IJSRST1841325 | Received : 20 Feb 2018 | Accepted : 28 Feb 2018 | January-February-2018 [(4) 2: 1514-1518]

II. RELATED WORK

Evaluators usually issue their own assessments as a part of the on-going learning method, and there’s a growing literature on however best to integrate and formalize these processes. Human effort required for the assessment is very high. It depends on several factors such as knowledge of the teacher, application level understanding of the teacher, criteria of the marking and time allotted. However it consumes very costly efforts and take huge time for the completion of the complete evaluation, verification and publishing of the result process.

Existing methods of computer based evaluation systems, often does not scale well and also don’t fully support features like: evaluation of subjective questions, offline examinations and delivery of dynamic content. These features are extremely desirable for evaluation. There is a need for alternate ways of designing such applications. Mobile Agents are an effective paradigm for distributed application [11].

Many architectures and features have been proposed for descriptive answer evaluation. The approaches are mainly based on feature or keyword matching, sequence matching and quantitative analysis. But the semantic or meaning wise analysis of descriptive answer is still an open problem. Considering the general structure of text analysis in natural language processing, most of the work has been done for syntactic analysis, but semantic, discourse and pragmatic are still being explored [12].

In 1960’s research on automated essay scoring system started, after 40 years of research four mature AES systems are commercially available. These systems uses large number of essay features and machine learning algorithms to predict and grade essays. In 1966 the first automated essay scoring system, Project Essay Grading (PEG), is developed by Ellis Page. It is widely used in testing companies, universities, and public schools. The PEG system uses shallow text features of essays and multiple linear regression to learn the scoring function. But scores given by the PEG system and scores given by human raters is proved to be high [6].This is because, PEG system only considers shallow text features of essay while ignores the essay content, leading that it is easy to be cheated by students. In the last of 1990s two automated essay scoring system are developed, one of them is Intelligent Essay Assessor (IEA), is developed. The IEA systems scores essay by measuring the semantic features and computes the semantic matrix of scored essays and unscored essays by a semantic text analysis method named Latent Semantic Analysis (LSA) [6]. Another one is E-rater, developed by Educational Testing Services (ETS) in America, and has been currently used for essay scoring in the Graduate Management Admissions Test (GMAT). In 2003 another mature commercial automated essay scoring system is developed based on artificial intelligence, IntelliMetric by Vantage Learning Company. These system extracts more than 300 text features, including both shallow text features and deep text features, feature extraction is complicated [6][13].

Jamsheedh C. V et al [5] have explained a paper for answer paper evaluation by using NLP tools. Here, features are extracted from scored sample dataset of answers and key. The extracted features are represented by bag of words model. The model is used to calculate the score of unscored answers. It evaluates scores for the exact sentence found in the bag of words only. It evaluates marks for the same type of answer more than once and the number of incorrectly classified instances are larger than the correctly classified instances.

Charusheela Nehete et al [1] describe an online assessment system that stores the answers after the assessment for processing. Due to the large execution time, it is not able to handle very complex sentence structures.

Ranjit Biswas [7] has explained answer paper evaluation based on fuzzy sets known as Fuzzy Evaluation Method (FEM). The paper compares the traditional evaluation approaches and automated grading approaches. Fem is a computer based fuzzy approach where a vector valued marking is used. The paper proposes a new generalized method of fem, gfem in which a matrix valued marking is adopted. Philip E. Robinson et al [8] have presented an online learning platform for teaching, learning, and assessment of programming. The open source online learning platform aids in teaching and assessment of computer programming in large classes. The paper describes the technology and implementation of the learning platform and new methods for automated assessment of programming assignments and exams.

III. PROPOSED SYSTEM

Many schemes and methods are currently available for evaluation of essays. But automatic evaluation and grading cannot been adopted for descriptive answers. In this approach, a novel method for automatic assessment of descriptive text answers is proposed.

In this assessment system, there are mainly three modules. The first part is the scanning phase. The scanning phase scans the document and identifies and extracts student answers. The created dataset consists of identified answers and its human evaluated score. The answers is given as input to the pre-processing phase which is the Natural Language Processing (NLP) part that enables to filter out the essential required parts from the answer sheets. The second part is the learning part which consists of training and testing. After the creation of the trained model, user of the system are able to provide the unscored answers in the form of pdf file to award marks. Training deals with creating a model by learning knowledge from the scored answers dataset and the answer key. Testing deals with the scoring of unscored answers based on the learned data in the trained model. The answer paper evaluation system makes the system capable of efficiently evaluating answer papers based on an answer key and a sample data set of scored answers with reduced human effort and reduced cost in a very short span of time under the right supervision of an evaluator. The proposed system accepts answer paper pdf files. Then, the relevant feature are extracted from the received answers and the created model is used to evaluate the marks of the descriptive answers.

The proposed assessment system extracts the semantics to efficiently represent the text in answers. A model is developed from answer key as well as scored answers to grade answers. The figure 1 shows the basic architecture of the proposed system.

The proposed system consists of Preprocessing, Comparison and Scoring. The first part of learning is the training which is used to create the trained model. During training the scored answers and high weightage given answer key is taken as the input for the training phase. The preprocessing extracts important features of the input. Then, each preprocessed answer and key is mapped into a vector space based on TF_IDF score and cosine similarity score between the produced word vector in the answer and word vectors in the key is calculated.

The testing deals with the scoring of unscored answers based on the learned data in the trained model. The unscored answers are converted into the TF_IDF vectors and cosine based similarity matching is performed based on the trained model. The scores corresponding to the most similar sentences to the sentences in the unscored answers are further used for grading

This stage extracts the relevant features of the text documents. The text document may be a key or answers. Natural Language Processing is used to extract the semantics of the answer sheets as a pre-processing step. The output of the pre-processing step is a set of unique words corresponding to each sentence in the answer. NLTK is a Natural Language Processing Toolkit that has a wide collection of predefined libraries for Natural Language Processing. Following are the key steps used in the proposed system of text evaluation in digitized descriptive answer.

The preprocessing consist of grammar check, tokenization, stop words removal, synonym and antonym checking, stemming. The first step is the tokenization where the process will convert a sentence into a set of words. After that we can ignore some of the commonly used words, the stop words such as is, was and so on. Stop words are words that does not convey any meaning. A text file "stop-words.txt" specifies the set of stop words is loaded into the program. After loading, it is compared to the words in the list. If a match is found, the word in the given list is removed.

The next step is antonym checking. The student may write answers with sentences that have negative meaning. Those negated words should be identified and can be expanded and added to the set of words. For extracting the antonyms, WordNet can be used. The next step towards preprocessing is stemming. Words with same meaning appear in various morphological forms. To capture their similarity, they are normalized into a common root form, the stem. This process is known as stemming. Example: For 'writing', 'wrote' and 'written', the stem is 'write'. Here we have used porter stemmer algorithm. Now, the duplicate words should be removed from the list in order to get a set of unique words corresponding to a sentence

For getting the vector representation of the sentence, we use TF-IDF calculation. Term frequency–inverse document frequency is a numerical statistic that is used to find out the importance of a word to a document in a collection or corpus. The number of times a term, t occurs in a document, d is called its term frequency. IDF measures the amount of information provided by the word. It is obtained by dividing the total number of documents by the number of documents containing the term.

A cosine based similarity checking algorithm is used to check the sentence similarity. A bag of words model is created based on similarity score between each key and every short answers. During testing the unscored answers are converted to its vector representation and the bag of words model is used to find the most similar sentence vectors with its score. The extracted marks of the most similar sentences are added to get the final score of the answer.

$$\mathrm{Sim}\,(s1, s2) = {(\overrightarrow{s1}\,.\,\overrightarrow{s2})}\Big/{\|\overrightarrow{s1}\|\|\overrightarrow{s2}\|} \qquad (1)$$

From the equation 1, $\overrightarrow{s1}$ and $\overrightarrow{s2}$ are the vector representation of the sentences. Cosine Similarity measure is used to find the similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. Based on the similarity score, the system will calculate the marks of the unscored answers.

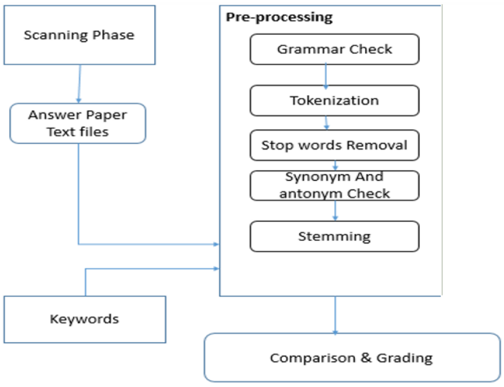

Figure 1. Proposed System Architecture. The architecture include Preprocessing Module, word to vector conversion and comparison and grading module. .

IV. CONCLUSION

This paper presents an approach for the design and implementation of automated handwritten short answer evaluation system. The proposed system works based on machine learning. It trains a model from the scored short answer paper dataset and a high weightage given key. The model is used to test the unscored short answer papers. The future scope in this project is that we can reduce the computation time by introducing hashing techniques into this system and extending the system for diagram evaluations.

V. REFERENCES

[1]. Charusheela Nehete, Vasant Powar, Shivam Upadhyay and Jitesh Wadhwani, "Checkpoint – An Online Descriptive Answers Grading Tool," IJARCS, April 2017.

[2]. J. Talwar, S. Ranjani, Anwaya Aras, and M. Bedekar, Intelligent Classroom System for Qualitative Analysis of Students' Conceptual Understanding", IEEE, 2013.

[3]. Jannat Talwar, Shree Ranjani and Anwaya, "Intelligent Classroom System for Qualitative Analysis of Students' Conceptual Understanding," IEEE Int. Conf. on Emerging Trends in Engg. And Technology, pp. 105–125, vol. 28, 2013.

[4]. Buddhiprabha Erabadda, Surangika Ranathunga and Gihan Dias, "Computer Aided Evaluation of Multi-Step Answers to Algebra Questions," IEEE Int. Conf. on Adv. Learning Technologies, pp. 45–65, vol. 28, 2016.

[5]. Jamsheedh C. V, Aby Abahai T and Surekha Mariam Varghese, "A Fair Assessment System For Evaluation And Grading Of Text In Degitized Descriptive Answer Scripts," unpublished.

[6]. Semire DIKLI and Tallahassee, "Automated Essay Scoring," Turkish Online Journal of Distance Education-TOJDE, Vol. 7, No. 1, Article: 5, January 2006.

[7]. Ranjit Biswas, An application of fuzzy sets in students' evaluation," ELSEVIER Fuzzy Sets and Systems, vol.22, no. 1, pp. 229{236, 1995.

[8]. Philip E. Robinson and Johnson Carroll, "An Online Learning Platform for Teaching, Learning,and Assessment of Programming,"' IEEE Global Engineering Education Conference (EDUCON), vol. 3, no. 6, pp. 547-556, April 2017.

[9]. Pantulkar Sravanthi and Dr. B. Srinivasu, "Semantic Similarity between Sentences," International Research Journal of Engineering and Technology (IR-JET) vol. 2, pp. 156-161, 2017.

[10]. Ming Che Lee, JiaWei Chang, and Tung Cheng Hsieh, "A Grammar-Based Semantic Similarity Algorithm for Natural Language Sentences," Hindawi Publishing Corporation Scientic World Journal Vol. 2014, Article ID- 437162, pp. 1-1, 2014 .

[11]. Magdi Z. Rashad, Ahmed E. Hassan, Mahmoud A. Zaher and Mahmoud S. Kandil, " An Arabic, Web-based Exam Management System ", IJECS – IJENS International Journal of Computer Sciences and Electricals ,February 2010, Vol. 10 , No:01.

[12]. Mohammad Salim Ahmed, Fahad Bin Muhaya, Lathifur khan and Sourabh Jain, " Predicted Probability Enhancement for Multilabel Text Classification using Class-Label Pair Associations," IEEE Conference on Evolving and Adaptive Intelligent Systems, April 2013, pp. 70-77.

[13]. Panchami K.S, Surekha Mariam Vargheseb and Aby Abhahai T, " Grading of Diagrams in Answer Scripts Using Support Vector Machine," International Journal of Control Theory and Applications, Vol.10,No.29,pp. 345-350, 2017