An Intelligent System for Evaluation of Descriptive Answers

submitted in partial fulfillment of the requirement
for the award of the Degree of

Bachelor of Technology
in
Computer Engineering

by

Manasi Beldar (2017130010)
Vinal Bagaria (2017130007)
Mohit Badve (2017130006)

under the guidance of

Prof. Sunil Ghane

Department of Computer Engineering
Bharatiya Vidya Bhavan's
Sardar Patel Institute of Technology
(Autonomous Institute Affiliated to Mumbai University)
Munshi Nagar, Andheri-West, Mumbai-400058
August 2020

Approval Certificate

This is to certify that the Project entitled "An Intelligent System for Evaluation of Descriptive Answers" by Manasi Beldar, Vinal Bagaria and Mohit Badve is approved for the partial fulfillment of Major Project-I for Final Year Project towards obtaining the Bachelor of Technology Degree in Computer Engineering from University of Mumbai.

Project Supervisor

Name: Prof. Sunil Ghane

Signature

External Examiner

Name:

Signature

Head of Department

Date:

Principal

Contents

List of Figures

1 Abstract

Examination is an efficient and a conventional way to test the knowledge gained by the students. The evaluation of examination is tedious and time consuming task. Automatic answer script evaluation makes this task convenient for teachers as it reduces the efforts and time taken. The existing systems provide teachers the facility to conduct automatic MCQ and short answers exam but evaluating descriptive answers automatically is a challenge.

To solve this problem, we propose an intelligent system which uses machine learning approach to automate the process. It considers question type, keywords, grammar, language, similarity with respect to the expected answer using cosine, jaccard, etc. similarity metrics and apply ensemble learning to predict the marks. The system also generates a personalized feedback after every exam for every question deriving the area in which the student is lagging.

2 Introduction

Evaluation of an examination is a tedious but an essential task for teachers. Students get their feedback about the subject knowledge they have earned through assessment of examination. After examination, the teachers spend most of their time evaluating the marks of the students and the evaluation. It takes bulk usage of human effort, time and cost. The manual evaluation involves external factors like the mood of the teacher, teacher-student relationship which lead to biasness. Motivation behind automation of answer-script evaluation includes fast processing, less manpower, independence of change in psychology of human evaluator, ease in record keeping and extraction.

The existing automated assesment systems work well only for MCQs or one-word answers. Automatic evaluation of objective answers is comparatively easy and well supported in many systems but, in case of descriptive answers, it is an open problem. The descriptive answer of the same question varies from student to student. Our system gives an efficient approach for the automation of the evaluation process of descriptive answers and also provide personalized feedback to the student about key areas in which he/she should improve.

3 Literature Survey

In the paper [1], NLP is used in conjunction with Deep Learning. Firstly, the answer of a student is converted into its GloVe vectors and those vectors are fed to the LSTM RNN model. The model gives semantic representation of the answer as an output which is then fed to the output layer for predicting marks on the basis of a trained network.

According to the paper [2], the teacher needs to input questions and answers along with keywords required and minimum no.of keywords in the expected answer. The keywords are checked directly if they are present in the given answer or not. The scoring is done by using 'edit distance' as a similarity metric.

In the paper [3], the answer is preprocessed using NLP. The vectors are created based on language dependent features such as noun, adjective, verb and their synsets using TF/ IDF. Marks are given by comparing these vectors with the expected answer vectors.

The paper [4] discusses the approach of forming vectors from the keywords are extracted from the answer. The 0/1 vectors are formed from these keywords by comparing with the keywords in the expected answer. Cosine similarity and term frequency methods are used for checking the similarity.

After summarizing the answer with the help of NLP, the concept map of the expected answer is generated. This map is stored for comparing the concept maps of different answers. The metrics involve usage of hierarchy, concepts and WordNet as a resource. This approach is followed by the paper [5].

The paper [6] uses the average of marks assigned by all the similarity metrics like cosine, jaccard, etc. The input answer is first preprocessed by keyword based summarization followed by pruning and stemming.

4 Research Gap Identified

There are many ways of representing the answer in a vector format. In one of the papers, vector representation of an answer was created using language dependent features like noun–verb–adjective–adverb and noun–verb metrics. This is useful only for literature subjects and may not work for theory subjects.

Different similarity metrics have different aspects of computing the similarity between answers. Some focus on structure and syntax (bi-gram, n-gram), some on the actual words (jaccard) and some on the frequency of words (cosine). Each metric has its own advantages and disadvantages hence, there is no single metric which individually can give good results for a generic system.

Existing systems rarely check synonyms while checking the similarity with keywords. The use of synonyms in the answer may be accepted in some cases. All the existing systems have only focused on answers for automatic evaluation but the evaluation should be according to the question as well. Questions having descriptive answers belong to different categories like factual, inductive and analytical. So these questions need to be treated differently as using the same method will increase inaccuracy.

To summarize, the main issues that need to be addressed are, the consideration of question type while evaluating an answer, the use of synonyms after pruning and stemming of a summarized answer and application of appropriate ML model to predict the marks.

5 Research Problems

The problems that we encountered are listed down.

1. Some of the theory subjects can have questions which do not require keyword specific answers. So, student can also write synonym of the keyword expected. The existing systems do not consider question type and synonyms while evaluating answers.

2. The existing systems use only one machine learning model or similarity metric for assigning marks. Sticking to only one model has its own benefits and drawbacks [7]. Considering different models and weighted similarity metrics can increase the reliability [2].

6 Problem Statement

Our automatic answer-script evaluation system requires sample test papers as input from the teacher. The questions will be classified according to their intent and the answers will be preprocessed and summarized with the help of NLP. The correctness of the answers will be checked taking into consideration various language and knowledge parameters. The marks will be predicted by the classification model. After evaluation, the student will get feedback to improve in areas, which he/she may be lagging in, like concepts, grammar, particular course outcome, etc. This will help the teacher in analyzing the class result collectively and focus on teaching the key areas in which the majority of the class is not performing well. It will also help to track a student's progress and analyze his/her improvement. The learning behaviour of a student will be predicted based on the progress report generated.

7 Motivation

The motivations behind implementing this automated system of answer-script evaluation are as follows:

- There is a lack of subjective mock tests because of unavailability of teachers for the correction of exam papers.
- The existing systems mainly focus on objective answers evaluation so there is a need for a similar system for descriptive answers as well.
- There is no consideration of synonyms in the current descriptive evaluation system and most of the time only keywords are being matched.
- Automatic evaluation systems mainly include less human interaction, quick processing and it also does not depend on change in psychology of human evaluators.

8 Objectives

Our key objectives for building this system of automatic answer-sheet evaluation are as follows:

- The system must offer subjective question paper evaluation with score and feedback for each question.
- The system must have accurate results with respect to short answers, long answers, grammatical mistakes, spelling mistakes.
- To provide a user-friendly and efficient system.
- To generate reports and analyze the performance of the students.
- To provide responsive design for ease of access by the students as well as teachers.

9 Expected Outcomes

The expected outcomes of our system are as follows:

- Our system should provide efficient evaluation of answers which takes into consideration various parameters like keywords, synonyms, antonyms, grammar, verbosity and structure of the answer.
- The gap between the scores obtained through manual evaluation and automatic evaluation through our system should be minimized.
- Our system should promote student-centric teaching for the progress of every category of student by analyzing the student progress and also by giving detailed feedback on their answers.
- Our system should assist the teachers in maintaining exam records and getting proper insights of student and class behaviours.

10 Scope

Our system will work for theory based mock examinations of subjects like artificial intelligence, machine learning, operating systems. Open ended questions are strictly not allowed as they may have completely varying answers from student to student, for example: why do you want to become an engineer? The answers should be strictly in textual format only. The system will not work for diagrams, flowcharts, formulae.

11 Requirements

Functional requirements define the basic system behaviour. They give us an idea of what response will be given by the system on giving particular input conditions.

Non-functional requirements give us an idea of how the response will be given by the system. They denote system related parameters like accuracy, performance, delay, etc.

11.1 Functional Requirements

- The system requires a sample answer paper from the teacher as an input. The input should be provided as a PDF file or a document.
- The whole answer sheet should get evaluated provided there are no open ended questions.
- After evaluation of the examination is done, the students should receive their feedback and progress reports. Also, class wise and course outcome wise analysis reports should be provided to the teacher.
- The progress reports of students should be automatically generated and sent to the respective student and teacher.

11.2 Non-Functional Requirements

- The sample answer paper should adhere to the predefined format and should not contain non-textual and open ended answers.
- The automatic evaluation definitely reduces the human efforts but it should also be time efficient and should be errorless evaluation.
- The feedback generated should give proper aspects of the areas where students need to focus on and reports given to the teachers should be easy to understand so that teachers can change the technique and way of teaching in a class.
- The progress report of the student should be saved and after enough data, it should give analysis to the student and teacher about the interests of the student and trends observed in the teaching-learning pattern.

12 System Design/ Methodology/ Algorithm

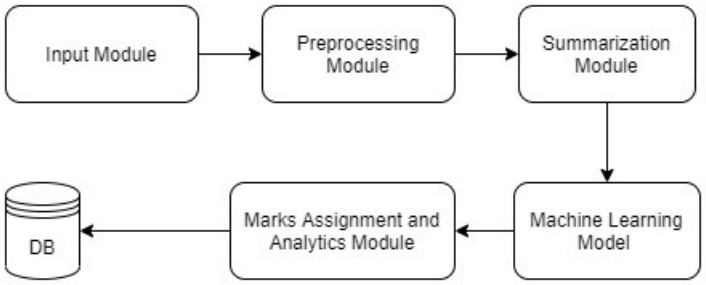

Figure 1: Architectural Diagram of Automatic Evaluation of Descriptive Answers.

Preprocessing will be done on the expected answer by removing stop words, punctuations and by applying lemmatization using WordNet, POS tagging, etc. The descriptive answer will then be summarized in order to extract keywords and only the important relevant sentences. This summarized expected answer will be stored in the database for further process. The same process of summarization will then be applied to the student's answer. The flow of the system is depicted in the Figure 2.

The keywords from the student's answer will be checked with the expected answer involving synonyms and the computed score will then be classified into different classes as shown in the Figure 3.

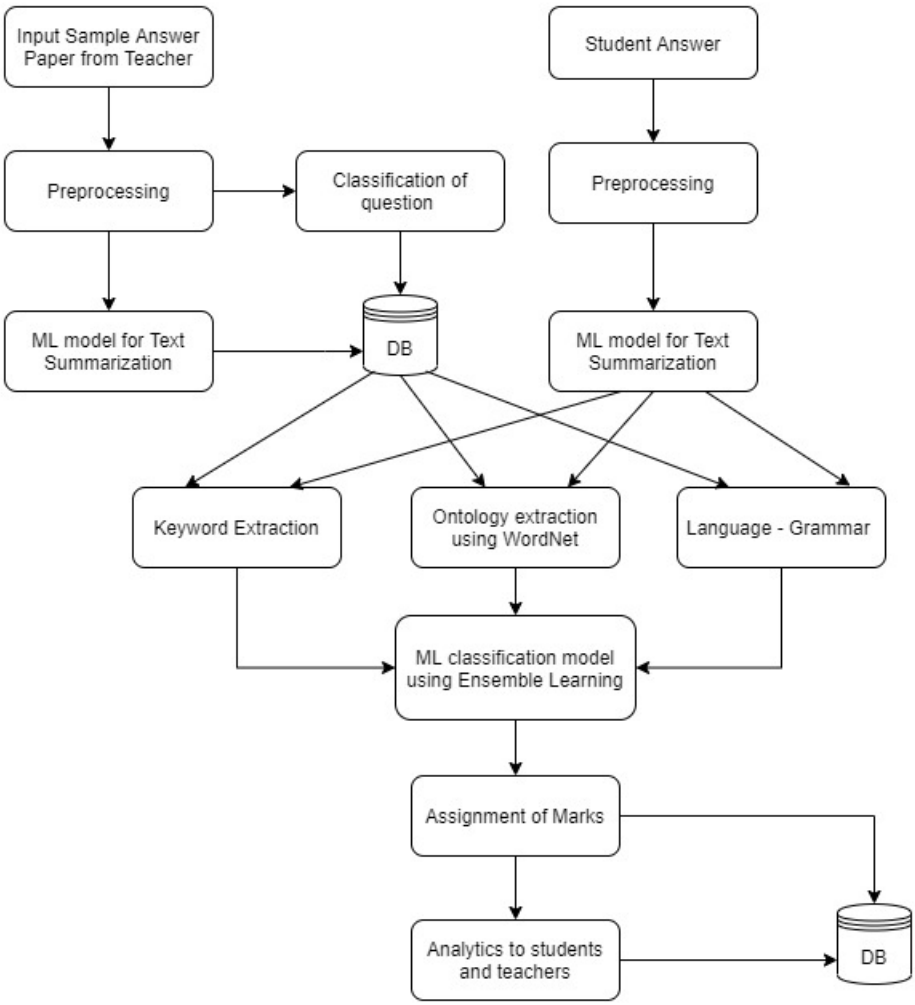

Figure 2: The Flow for Evaluation of Descriptive Answers.

Keywords Classes	Percentage Score
Excellent	> 90%
Very Good	75 - 90%
Good	61 - 74%
Ok	51 - 60%
Poor	35 - 50%
Very Poor	< 35%

Figure 3: Keyword Scores Mapping.

Then, the grammar and language related syntax will be checked and the corresponding class as depicted in figure 4 will be assigned to the grammar score of the student. We have kept less strict categorization in terms of grammar as knowledge is given more importance in theory subjects.

Grammar Classes	Percentage Score
Excellent	> 70%
Good	40 - 70%
Poor	< 40%

Figure 4: Grammar Scores Mapping.

Similarity Metrics will be applied to the student answer and expected answer. The score will be considered as the weighted cumulative score of the different metrics applied. The figure 5 represents the formula for the calculation.

$CSS = W1*S1 + W2*S2 + \ldots + Wn*Sn$

Where

CSS is the Cumulative Similarity Score

S1, S2, …, Sn are the individual scores of all similarities

W1, W2, …, Wn = Individual Weights

$W1 + W2 + \ldots + Wn = 1$

Figure 5: Cumulative Similarity Score Calculation.

The whole system is divided into different modules according to the flow as given in the figure 1. As we are using separate sections for language and knowledge, feedback will be given to the student in which area he/she is lagging. The section wise cumulative score and course outcome wise analysis will be shared with the teacher. Question wise, section wise, exam wise and course outcome wise progress report of each student, will be made available to the teacher and student himself/herself.

13 Contribution

Our system will mainly contribute to the more focused and personalized teaching-learning experience and also to evaluate mock exams automatically.

- The proposed system should overcome the problem of considering question type and synonyms while automatically evaluating the answers.
- The teachers may not be available for correcting each and every exam and hence won't be able to give personalized feedback. Our system will provide an efficient way to solve this problem for theory based mock exams.
- More and more exams can be taken as evaluation will reduce the efforts and time. Also, more focused teaching experience can be given to the students.

14 Conclusion

The proposed solution considers key parameters for evaluation namely the type of question, structural and conceptual aspects, various similarity metrics with appropriate weightage for a particular question. The system is user-friendly as the sample paper input could be in various formats and also the student feedback and progress analysis will be in the form of various charts and graphs for better understanding. It helps in eliminating the flaws in the existing manual evaluation system and promoting experience of personal learning environments. In future, the system can be moulded as a full-fledged automated distant learning exam conduction and assessment platform.

References

[1] Neethu George, Sijimol PJ, Surekha Mariam Varghese, "Grading Descriptive Answer Scripts Using Deep Learning", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-5 March, 2019.

[2] Md. Motiur Rahman, Ferdusee Akter, "An Automated Approach for Answer Script Evaluation Using Natural Language Processing", IJCSET(www.ijcset.net) — 2019 — Volume 9, 39-47.

[3] I. Das, B. Sharma, S. S. Rautaray and M. Pandey, "An Examination System Automation Using Natural Language Processing," 2019 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2019.

[4] S. P. Kar, J. K. Mandai and R. Chatterjee, "A comprehension based intelligent assessment architecture," 2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE), Hong Kong, 2017.

[5] S. B. Salem, L. Cheniti-Belcadhi and R. Braham, "A concept map based scenario for assessment of short and open answer questions," 2017 International Conference on Engineering & MIS (ICEMIS), Monastir, 2017.

[6] Ramamurthy, M., & Krishnamurthi, I. (2017). Design and Development of a Framework for an Automatic Answer Evaluation System Based on Similarity Measures, Journal of Intelligent Systems, 26(2), 243-262.

[7] A. Kaur and M. Sasikumar, "A comparative analysis of various approaches for automated assessment of descriptive answers," 2017 International Conference on Computational Intelligence in Data Science(ICCIDS), Chennai, 2017.