ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 10, October 2018

An Automatic System to Enhance Assessment Marking in Tertiary Education

Dr. Md. Waliur Rahman Miah[1], Amina Shaikh[3], Dr. Belal Chowdhury[3], Dr. Md. Nasim Akhtar[4]

Associate Professor, Department of CSE, DUET, Gazipur, Bangladesh[1]

M.Sc. Eng. Student, Department of CSE, DUET, Gazipur, Bangladesh[2]

School of IT & Engineering, MIT, Melbourne, Australia[3]

Professor, Department of CSE, DUET, Gazipur, Bangladesh[4]

ABSTRACT: In a tertiary level education assessments are powerful tools for a teacher to influence the learning of students. To be effective those assessments need to be accurately marked and comprehensive feedback should be given to the students within a short period of time. However manual marking is a painstaking chore for a teacher, especially when the class size is very large. The situation can be improved by automizing the marking task. In this paper we propose an automatic marking approach utilizing state of the art data mining and language processing techniques that can automatically generate deserving marks of narrative type answers provided by a student. Such a system can be advantageous to an academic marker (e.g., lecturer and tutor) as well as to a student examinee. In this paper, we propose an automatic marking system to provide easy marking and comprehensive feedback in shortest time by the academic markers. Obtaining quick feedback students can identify and understand the areas of weakness and address them immediately in order to improve their learning.

KEYWORDS: automatic marking, data mining, natural language processing, document similarity, intelligent assessment.

I. INTRODUCTION

This paper is showcasing findings of a preliminary experiment that support our idea of ongoing research into an automatic marking scheme in tertiary education. The current research is motivated by the need for quick marking and providing effective feedback to students during their formative assessments (such as tutorials, assignments, and reports) throughout the academic term/semester. The proposed system can also be implemented, if proper resources are available, in the summative assessment at the end of the semester.

Expediting marking is critical for providing timely and accurate feedback to students and maintaining the integrity of results for a large cohort of students. They can self-monitor their academic progress by analyzing the feedback provided by their teacher on time. If the feedback is delayed until the end of the semester, there is no chance for the student to be evaluative and adjust their pace of learning. Therefore, educational research also emphasizes on such feedback to students [1], especially for ongoing formative assessments throughout the semester. Course evaluation surveys conducted in [2] and [3] demonstrate that students are dissatisfied with the amount and quality of feedback they receive. Burrows and D'Souza found that there is a significant demand on academics to provide timely, personalized and detailed feedback [4]. An inefficient manual approach of marking may not fulfil that demand, especially for large classes. The emergence of new online tools/technologies (such as an API from documentsimilarity.com, MapReduce/Hadoop, etc.) presents an exciting new opportunity to improve the way of academic marking processes. In addition, manual approach is also prone to error due to bias or exhaustion. In this current paper we try to address those problems by providing an automatic marking approach for university assessments.

The remaining of this paper is organized as follows:

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 10, October 2018

Section II provides a short literature review on related works. An overview of our proposed approach is illustrated in section III. Data collection, and experimental setup are explained in section IV. Also results are discussed in the same section. Finally, section V concludes our findings and outlines future directions.

II. RELATED WORK

Research in automatic marking has a long history dating back to the early work of Page [5]. Since then, automatic marking continually drawing interest in the research field.

Burrows et al in [6] provide an overview of the different techniques and methodologies used in different time for addressing the problem of automatic grading of short essay (ASAG). The authors' historical analysis identifies 35 different ASAG systems. Many of those are still subject to open research issues.

Burrows et al in [7] analyzed different kinds of marking tools for example: Moodle Workshops, TurnitinGradeMark, Waypoint and WebMark . A sample assignment was given, and the tutors were using all four systems to mark the sample assignment and evaluate the system according to some predefined criteria, for example: flexibility, features, functionality, usability, and navigation. Different systems were given priority over others in different features. However in overall ranking, Moodle was most applauded for non-functional requirements such as usability, layout, navigation and accessibility. The tools did not automatically marked the assignment,instead the tutor marked the assignment using those tools.

Biggam in [8] illustrates, the benefits of Automated Assessment Feedback for staff as well as students. A semi-automated assessment feedback system, using the simplest of technologies, is provided. A marking sheet was created in MS Word containing a marking table with different marking criteria or topics (eg. introduction, main body, conclusion, structure, reference etc.). A list of Feedback comments was written as macros for each topic. Markers could only choose an appropriate one from those list of comments for feedback on a particular marking topic. In this way a consistency of feedback was achieved among different markers. However the marks were given by human markers not by an automated system.

Very short type answers are marked using the system developed by Yorke et al in [9]. The system itself does not do the marking but clusters similar answers of different students. When the marker gives a mark to an answer of a student, the same mark is copied for all other students of the same cluster. To devise the clusters, term matching techniques are used; not semantics or language processing. Though a clustered approach significantly reduces the time spent on marking when the number of students is very high but certainly it is not an automated marking system.

Pearson Inc. uses a copyrighted and licensed system for automatic marking of Pearson Test of English -Academic (PTE-A) [10][11]. We searched but could not find any suitable paper that explains the complete methodology used in the automatic marking in PTE. The information we found from different papers fromPearon Inc., including a white paper [12], are that the Pearson Automatic Marking system uses Latent Semantic Analysis [13] along with Artificial Intelligence based natural language processing techniques. The marking system is trained with a very large data-base of 10,000 essays [14]. The results shown in those papers are very good, and the system is being used in the PTE test of English conducted by Pearson (PTE-A). However the exact methodology has yet not been disclosed, perhaps due to commercial and copyright issue. In contrast to the Pearson's commercial product, our current research aims to develop a system which is open to all, especially to academics and researchers.

III. AUTOMATIC MARKING MODEL IN TEACHING

A generic view of our proposed automatic marking approach is shown in Fig. 1. The heart of the system is the automatic marking engine. The examiner provides a sample model answer for each question which is kept in an internal database. A student's answer is the target input text to be marked by the system. The automatic marking engine compares the student's answer with the model answer and calculates the appropriate marks.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 10, October 2018

Different text comparing techniques can be the potential candidates to build the automatic marking engine. Measuring the correctness of grammatical syntax, semantic discourse, cosine similarity, WordNet semantic similarity, latent semantic similarity, ontology (knowledge) based similarity are some examples of different text comparing techniques. In this current paper, we present experiments conducted using a document similarity API from documentsimilarity.com [15]. According to that website, the similarity is measured using advanced Natural Language Processing and Machine Learning technologies that can be used to analyse the semantic similarity of two text documents. We keep the other techniques as a venture for future work.

IV. EXPERIMENT AND RESULT

In this section first we describe the data collection and experimental setup, then we analyze the result.

A. *Data Collection*

To collect our data, we took the students' answers to a question from a real class-test in Dhaka University of Engineering and Technology (DUET), Gazipur, Bangladesh. We considered each student's answer as a separate sample document. The exam and class tests in DUET are paper based. Therefore, to obtain the digital data the students' answers were typed by using notepad. There were 40 students responded to the concerned question, so after data-entry we obtained a data-set of 40 samples. This is the preliminary stage of an ongoing research. In this stage, we rather focused on visualizing our idea than going through a rigorous experiment with a big data-set. The small data set of 40 samples gives us some promising results that corroborate our idea and motivates us to advance in the proposed research direction.

B. *Experimental Setup*

Our experimental setup follows the proposed system described in section III. The course-teacher provides a model answer for the question. That answer is considered as the gold standard with which each answer of a student is to be compared. The comparison is accomplished inside the marking engine and an automated machine generated mark is given. For comparison, the marking engine uses a document similarity measure, based on cosine similarity and backed by natural language processing and machine learning. For convenience, we used an API from documentsimilarity.com to measure the similarity for now, however we plan to eventually build our own app for this purpose.

Let us assume the model answer provided by the course-teacher as M. Each student's answer is considered as Si. The similarity between M and Si is measured by the formula:

$$Sim(M,Si) = docSim(M,Si);$$

Where *docSim* is the document similarity. It is represented as a cosine similarity [16] ranged from 0 to 1, where 0 means absolutely different and 1 means absolutely same.

To compute the machine generated marks the similarity is multiplied by the allocated mark of the question. For example say the marks allocated is 10, and the *sim(M,Si)* is 0.6; then the mark is computed as $10 \times 0.6 = 6$.

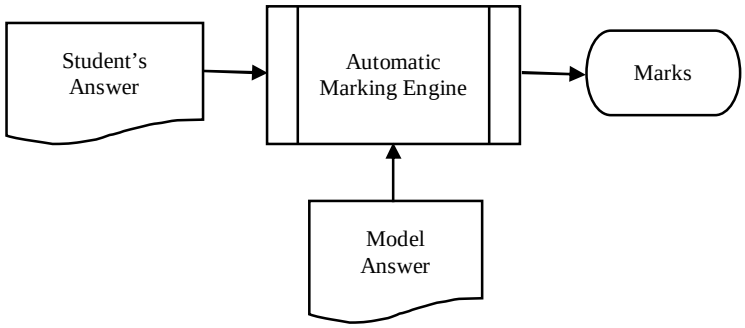

Fig. 1. Overview of the proposed automatic marking system.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 10, October 2018

To obtain human given marks, the students' answers were separately marked by two human-markers; one was the course-teacher himself and the other one was a post graduate student in computer science and engineering having a good knowledge on the course.

In the next section we present the mean absolute error and inter-rater-agreement as the evaluation metric of our proposed system.

C. *Result and Analysis*

Fig. 2 presents a column chart representing the Mean Absolute Errors (MAE) between each pair of markers. The notations H1, H2 and M represents human marker-1, human marker-2, and machine marker respectively. The average human marks are also calculated by averaging the two different marks given by the two human markers H1 and H2. The 'avg' represents the average mark. H1-H2 column is the MAE between two human markers H1 and H2. MAE between human marker H1 and machine marker (M) is given as H1-M. Comparison between human marker H2 and machine marker is given in H2-M. MAE between average marks and the machine-marks is presented in the column titled as Avg-M.

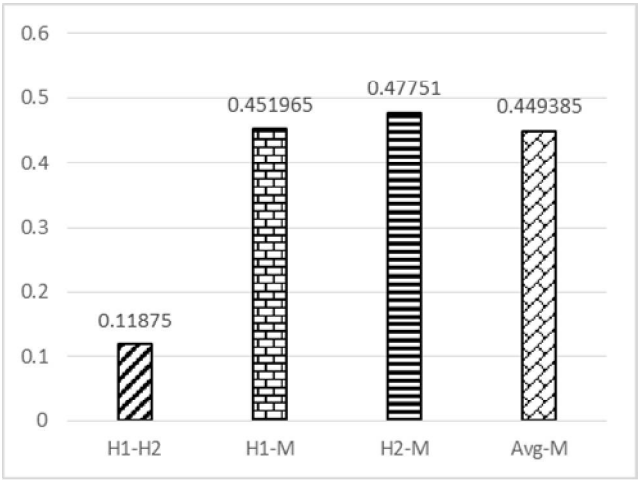

Fig. 2. Mean Absolute Errors (MAE) between each pair of markers.

The column charts in Fig.2 show that the mean absolute error between two human markers is 11.875%. Both the human markers are experts in the subject matter of the question. Therefore we consider both the marks given by those experts for an answer are correct. That is 12% (11.875 ≈ 12) variation of markscan be considered as correct marks.A variation of marks within a tolerable range is an accepted practice where different tutors are used to mark the same question. Therefore in our calculation of inter-rater-agreement, we consider it as an agreement between two markers even though the marks varies within the range of 12%.

The column chart in Fig. 3 presents Inter-Rater-Agreements (IRA) between each pair of markers. The notations H1, H2, M and Avg bear the same meaning as in Fig. 2, for example: the IRA between two human markers is presented in the H1-vs-H2 column, and the column titled Avg-vs-M shows the IRA between the average marks and the machine-marks.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 10, October 2018

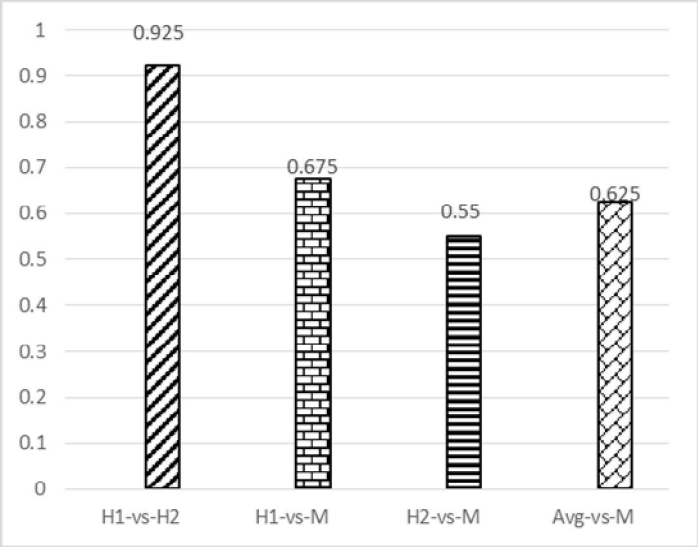

Fig. 3. Inter rater agreements between each pair of markers.

Analysing column charts in Fig. 3 we can see that the agreement between the two human markers is the highest with 92.5%. The very high agreement is obvious, because both of the two human markers are experts in the subject matter. They are intelligent and marked without any bias. However the small amount (7.5%) of disagreement between them shows that there are some cases where the human markers tend to disagree despite their expertise. Keeping this in mind we understand that there will be agreement as well as disagreement in the case of human versus machine marking. From the column charts we see that, though the amount of agreement between human-versus-machine markings are lower than human- versus -human marking, yet scored a fairly high agreements. The highest agreement between a human versus machine is with human marker H-1, achieving an IRA of 67.5%. In the case of machineversushuman marker H-2 the agreement is 55%, which is 12.5% lower than the IRA of H-1vs machine. When compared with the average of both humans given marks, the machine agreed on 62.5% of times.

V. CONCLUSION AND FUTURE DIRECTION

Analyzing the results provided in the previous section we can conclude that building an automated marking system using document similarity measure is not infeasible. The results show that our proposed automated marking approach (the machine-marker) gives a fairly high interrater agreement (IRA) with a human marker. However the amount of IRA is much lower than the amount of IRA between two human markers. To achieve a very high IRA with human, the marking engine requires a thorough investigation. Our proposed approach uses only the document similarity measure as the core of the marking engine. More sophisticated techniques, for example: measure of correct grammatical structures, discourse, context and concepts etc., were not used in the experiments of the current research. These can be avenues to explore in the future.

REFERENCES

1. R. Higgins, P.Hartley, and A.Skelton, “The conscientious consumer: Reconsidering the role of assessment feedback in student learning”. Studies in Higher Education, vol. 27, issue1, pp. 53-64. 2002. http://dx.doi.org/10.1080/03075070120099368.
2. Nicol, D. (2010). “From monologue to dialogue: Improving written feedback processes in mass higher education”. Assessment and Evaluation in Higher Education, vol. 35, issue5, pp. 501-517, 2010. http://dx.doi.org/10.1080/02602931003786559.
3. M. Price, K. Handley, J. Millar, and B. O'Donovan, “Feedback: All that effort, but what is the effect?”, Assessment and Evaluation in Higher Education, vol. 35, issue3, pp. 277-289, 2010. http://dx.doi.org/10.1080/02602930903541007.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 10, October 2018

4. S. Burrows, and D. D'Souza, "Management of teaching in a complex setting". In Proceedings of the Second Melbourne Computing Education Conventicle. pp. 1-8, Melbourne, Australia, 2005.
5. E.B.Page, (1966). "The imminence of grading essays by computer". Phi Delta Kappan, vol. 47,issue5, pp. 238–243, 1966.
6. S. Burrows, I. Gurevych, and B. Stein, "The Eras and Trends of Automatic Short Answer Grading", International Journal of Artificial Intelligence in Education, March 2015, vol. 25,issue 1, pp. 60–117, Springer, New York, 2015.
7. S.Burrows, andM. Shortis, "An evaluation of semi-automated, collaborative marking and feedback systems: Academic staff perspectives". Australasian Journal of Educational Technology, vol. 27,issue 7, pp. 1135-1154,2011.
8. J. Biggam. "Using Automated Assessment Feedback to Enhance the Quality of Student Learning in Universities: A Case Study". In: Lytras M.D. et al. (eds) Technology Enhanced Learning. Quality of Teaching and Educational Reform. Communications in Computer and Information Science, vol. 73. pp. 188-194, Springer, Berlin, Heidelberg, 2010.
9. J. Yorke,W. Gibson, H. Wilkinson, "Towards sustainable marking practices and improved quality of feedback in short – Answer assessments". In: ATN assessment conference, University of Technology Sydney, Sydney, NSW, Australia, 2010.
10. Pearson, "Pearson Test of English-Academic", 2018. [Online]. Available:https://pearsonpte.com [Accessed: March, 2018].
11. Pearson Education,"Intelligent Essay Assessor (IEA) Fact Sheet", 2010. [Online]. Available: https://images.pearsonassessments.com/images/assets/kt/download/IEA-FactSheet-20100401.pdf [Accessed: March, 2018].
12. Pearson Inc., "Pearson's Automated Scoring of Writing, Speaking, and Mathematics (White Paper)", 2011. [Online]. Available: https://images.pearsonassessments.com/images/tmrs/PearsonsAutomatedScoringofWritingSpeakingandMathematics.pdf [Accessed: March, 2018].
13. T.K. Landauer, P.W. Foltz, and D. Laham,"Introduction to Latent Semantic Analysis". Discourse Processes, vol. 25, pp. 259–284, 1998.
14. Pearson , "Pearson Test of English Academic: Automated Scoring". Pearson Education Ltd, 2011. [Online]. Available: https://pearsonpte.com/wp-content/uploads/2015/05/7.-PTEA_Automated_Scoring.pdf [Accessed: March, 2018].
15. Documentsimilarity.com, 2018. [Online]. Available: http://documentsimilarity.com/document-similarity-demo[Accessed: March, 2018].
16. C. D. Manning, P. Raghavan, and H. Schütze, "An Introduction to Information Retrieval", Cambridge University Press, Cambridge, England, 2009.