Evaluation of Structured Questions Using Modified BLEU Algorithm

Fadzilah Siraj
Big Data Task Force, School of Computing, College of Arts and Sciences
University Utara Malaysia, Sintok, Kedah, Malaysia

Mohamed Salem Ali
School of Computing, College of Arts and Sciences
University Utara Malaysia, Sintok, Kedah, Malaysia

Abstract - Abstract- This paper describes and exemplifies an application of a Structured Exam Questions Test Bank and Evaluation Using Modified Bilingual Evaluation Understudy (BLEU) Algorithm, a software system developed in pursuit of robust computerized marking of free-text answers to open-ended questions. It employs the Information System Development Research Methodology with modified BLEU Algorithm and Expert System for similar words. The system was developed to facilitate in managing and administrating structured questions for client/server architecture based on intranet. The system incorporates a number of processing modules specifically aim at providing an automated marking to reduce spelling errors, calculating scores, managing and administrating an exam. The system was trial-run by a group of students and lecturers, and modifications particularly on the interface have been modified and implemented. Problems and limitations discovered were then discussed and recommendations made to overcome the limitations for the future development of the research.

Keywords – Structured Exam Questions, Evaluation, Modified BLEU, Expert System

I. INTRODUCTION

Assessing students' answers is a very time-consuming activity that makes teachers cut down the time they can devote to other duties. Assessment is considered to play a central role in the educational process, Most of the existing education courses rely on objective testing exercises, such as Multiple Choice Question (MCQ) or fill-in-the-blank items. However, in order to fully assess the students' learning progress, these should be complemented with open-ended questions [1]. It has been noted that assessment based only in multiple -choice, fill-in-the-blank or yes/no questions is not accurate enough to measure the amount of knowledge the students have acquired, or whether they have understood the subject, Therefore, the field called Computer-Assisted Assessment (CAA) of open-ended questions has been created to study how the computer can be used to automatically assess students' free-text answers. However, evaluating free-text answers automatically is not a trivial task to tackle, as it requires some degree of natural-language understanding of the students' answers. A classification of these techniques with examples of existing systems that use them is given in [2]. The original BLEU Algorithm was introduced by [3] for evaluating the performance of machine translation system. In 2005, [4] proposed a modified BLEU Algorithm by not penalizing brevity and improved performance when compared with the unmodified BLEU Algorithm and a dump baseline.

The most highly valued activity for a lecturer is to teach but assessing students' answers is a very time-consuming activity that makes lecturer reduces the time they can devote to other duties. In some cases, they may even have to reduce the number of assignments given to their students due to lack of time for marking. In other case, when the number of student increases, the marking load also increases.

In reality, an exam question comprises of multiple choices questions and/or essays questions. To mark multiple choices exam is an easy task. However, marking structured and essays question consume more time than marking multiple choices questions. System for marking multiple choices exams are quite easy to develop, however marking structured or essay question can be very difficult to develop. Since not many systems are available to cater for structured/essay exam question, this research attempts to build such a system based on modified BLEU algorithm. Specifically, this study aims to automate text evaluation using modified BLEU algorithm for structured questions and answers. In addition, the study also aims to integrate modified BLEU algorithm that incorporates expert system into text evaluation engine.

The rest of the paper is organized as follows: Section II covers related work in text evaluation. The methodology of the study is presented in Section III. Results of the study is described in Section IV, followed by the conclusion of the study (Section V).

II. AUTOMATED TEXT EVALUATION

Several automated text evaluations have been used in evaluating free text answers with increasingly better results such as Information Extraction (IE), Natural Language Processing (NLP), Statistical Lexical Relationships (SLR), Anaphora Resolution (AR), Latent Semantic Analysis (LSA), Text Categorization Techniques (TCT) and Bilingual Evaluation Understudy Algorithm (BLEU). The pioneer in the field of CAA of free text answers was [5] with the Project Essay Grader (PEG) aims to improve the assessment process. The Intelligent Essay Assessor (IEA) developed at the Colorado University in USA primarily focuses on the content based on Latent Semantic Analysis (LSA) [6]. Another system known as Intelli Metric was developed by the American company Vantage Learning Technologies used an Artificial Intelligence approach to assess both the style and the content of free text answers [7]. To overcome the weaknesses of E-rater [8] the paraphrase recognizer C-rater was developed [9]. Since 1999, Educational Testing Service (ETS) from United States started to use E-rater in the GMAT exam. Two years later, ETS started the Criterion project [10] whose back-end is E-rater.

Another research lines were those initiated by [11] whose system was based on text categorization techniques (TCT), text complexity features, and linear regression methods. For example, IEMS was developed based on the Indextron technique [12] while SEAR (Schema Extract Analyze and Report) provides automated marking of the essay content and essay that is based on pattern-matching techniques [13]. IEMS was developed by [14], based on the Indextron technique. [15] created Apex Assessor that similar to IEA [6] is also underpinned by LSA.

In 2001, [16] from the University of Portsmouth (UK) followed another research line that led them to the Automated Text Marker (ATM). ATM looks for concepts in the text, and their dependencies, to finally give two independent scores, one for the style and another for the content. Automark was also developed by [17] that employs NLP techniques to perform an intelligent search of free text responses according to predefined computerized mark scheme answers. Another point of view is the given by [18] that created their Bayesian Essay Test Scoring SYstem (BETSY based on statistical analyses.

[19] and [20] developed the TANGOW (Task-based Adaptive learner Guidance on the Web) system, which supports the specification and dynamic generation of adaptive web-based courses, so that the course components are tailored to each student at runtime. It is based on n-gram co-occurrence metrics which allow the system to perform a vocabulary analysis and to study how similar student and teacher answers are. In 2002, the Paperless school free text marking engine (PS-ME) appeared, developed by [21], in which all the Bloom's taxonomy competencies [22] are covered and the answers are processed using NLP techniques. In 2003, two new systems came into view: Auto-marking, created by [23] is based on NLP techniques and pattern matching; and CarmelTC, proposed by [24] assesses students's writing with machine learning classification methods and a naive Bayesian classification.

Automated writing evaluation (AWAE) appears to complement instructor's input to provide immediate scoring and qualitative feedback about students' writing development or answers. When integrated with Criterion software developed by ETS, the findings reveal that both instructors and students gain advantage from AWE. The Criterion enables the instructors to determine students' writing needs while students revise their papers to obtain higher score. Automated Essay Evaluation (AEE) and Automated Essay Scoring (AES) systems [25] are being increasingly adopted in the United States to support writing instructions [26]. Earlier related research [27] [28] indicates that users increase their writing motivation. In line with Wilson and Czik's findings, AEE enable the teachers to focus on higher-level writing skills, while increasing student's writing motivation and writing quality.

It is important to highlight that the aim of the research is not to get instructors to do less work, but to get students to write. In fact, students are given the possibility of being assessed more times (the computer does not get tired) and of receiving a more detailed feedback (the computer can perform more complex analyses than any human being), without trying to substitute teachers' skills such as evaluating general personal opinions or creative writing, and giving always the teachers' criteria the maximum priority [2] [29]. [30] incorporates ROUGE to determine the correlation between the manual and automated evaluations of care episodes from electronic health record. There is high correlation between the manual and automated evaluations suggests that the less labour-intensive automated evaluations can be used as a proxy for human evaluations when developing summarization methods.

In the field of CAA, automated processing of free-text material received from students is becoming a necessity. The range of such material may run from single sentences to whole essays. Even as seemingly small a problem as student answers to open-ended questions poses a variety of serious Natural Language Processing (NLP) challenges. It calls for different approaches, depending on the didactic purpose of the exercises. This, in turn, affects the nature of the textual material that can be submitted to automated assessment [31] [29] [25].

III. METHODOLOGY

In this research, the general research methodology is adopted from [32]). In the development phase, was implemented the expert system was incorporated with BLEU Algorithm in dealing with similar words that occur in the exam question and the answer. The system development life cycle (SDLC) is an organized set activities that guide those involved through the development of an information system shown in Figure 1 illustrates the steps using ISDRM for creating Structures Exam Question Test Bank and Evaluation.

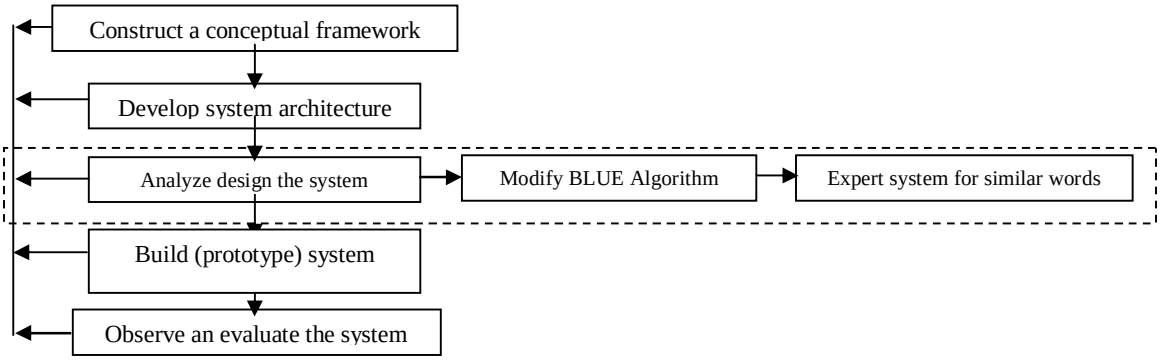


Figure 1. Information System Development Research Methodology

The system architecture for the Student and Lecturers are illustrated in Figure 2 and Figure 3. For student's answers evaluation, the modified BLEU Algorithm is utilized and this algorithm is described in detail in the following subsection.

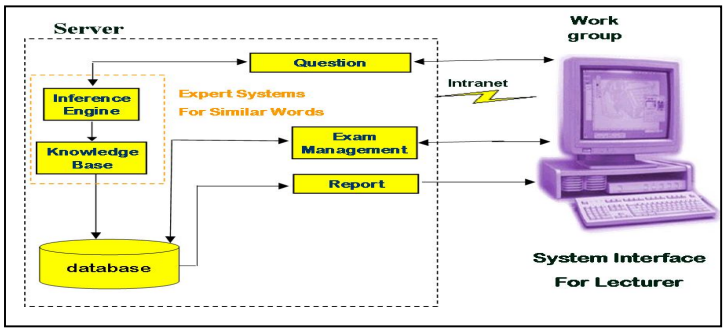


Figure 2. System Architecture for Lecturers

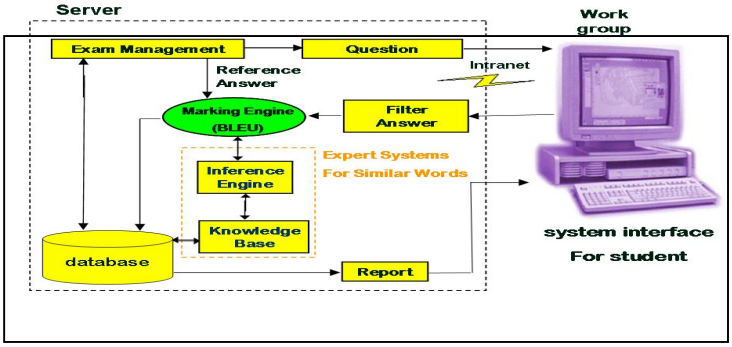


Figure 3. System Architecture for Student

From [3], the original BLEU algorithm is written as in equation (1).

$$s_i = \frac{c_{test,ref}}{c_{test}} \quad (1)$$

A weighted geometric mean is given by equation (2).

$$s_N = e^{\sum_{i=1}^{N} w_i \log(s_i)} \quad (2)$$

Where the weight w_i is typically kept constant for all i (w_i=1/N for all i).
The brevity penalty is given by equation (3).

$$b = \begin{cases} e^{(1-r/c)} & \text{if } t < r \\ 1 & \text{if } t > r \end{cases} \quad (3)$$

Hence, the overall score can be written as equation (4).

$$s_{bleu} = b s_N \quad (4)$$

Following [4], the modified BLEU Algorithm can be written in the Procedure as follows:
si = si + (1 / n)
*marke_student_q = mark_te * si*
total_te = mark_te + total_te
total_st = total_st + marke_student_q

In addition, Expert system is also integrated with BLEU Algorithm to cater for similar processing.

A. Expert System for Similar Wards

In the reference answer, the lecturers use specific words, however the student may use different words that have the same meaning. To cater for this problem, a Rule Based System for Similar Words is consulted for similar word. The probability allocated by the expert to similar words ranging from 0.1 to 1.00. Figure 4 shows the system Architecture for similar words.

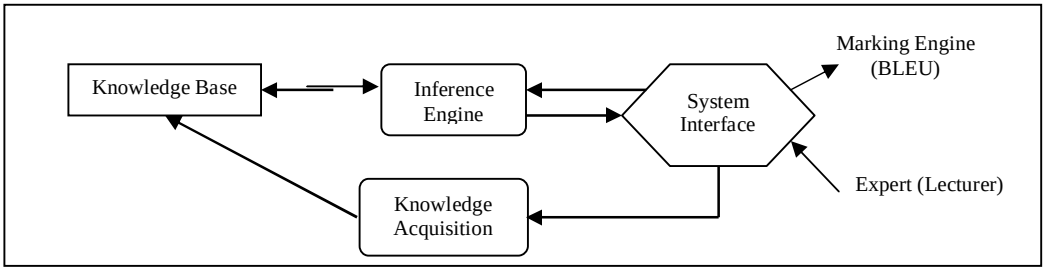


Figure 4. Expert System for Similar Words Architecture

B. Knowledge Base

The knowledge base of the system contains ten parts of the probabilities values that correspond with the stored synonym for each word in the reference answers. For demonstration purposes, the Knowledge Base implementation for similar words is shown in Figure 5.

	Word	100%	90%	80%	70%	60%	50%	40%	30%	20%	10%
	ability										
	mimic										
	function										
*											

Figure 5. An Example of Knowledge Base implementation

C. Inference Engine

The engine obtained is processed through simple filters that eliminate common words, prepositions, conjunctions and the like. Next, the engine determines key words by successively taking words in the sentence and Searching for it in the keywords and synonyms database. If a word is found in the database, the search is continued in case another one exists in the reference answer. Otherwise, a new word is added to the system by entering a learning mode which allows the lecturers to enter synonyms to existing database.

In assessing student's answer, the engine compares keywords in the sentence of student answer with key words in reference answer and synonyms available in the database. Hence, if n is the synonym of the word m, which is input from the Student answer, n is checked against a list of synonyms. Each synonym has percentage values between 0.1 to 1 (inclusive) to show the similarities between word and synonyms. The coding for implementing Procedure for Inference Engine is shown below.

Procedure for Inference Engine:

```
st = RemoveExtraSpaces(Trim(Reference_Answer))   /For removed extra spaces from answer
   Reference1() = Split(st, " ")                 /For cut the answer to keyword
  For i = 0 To UBound(Reference1)                /Put each keyword into combo like array
  Combo5.AddItem Reference1(i)
   Next i
For i = 0 To Combo5.ListCount – 1     /For filter each keyword by removed closed-class
Combo5.ListIndex = I                  / words and naïve Word
test_filter = False
filter_finel (Combo5.Text)
Next
For i = 0 To Combo6.ListCount - 1
Combo6.ListIndex = i
s3 = "select * from synonyms"       / For searching for keyword in database (Knowledge Base)
s3 = s3 & " where Original_Word='" & Trim(key_word) & "'"
Next
```

D. User Interface

The user interface allows the Lecturer to enter synonyms for new word, the table is then displayed the probability allocated by the expert to similar words ranging from 0.1 to 1.00. Figure 6 shows Expert System interface for similar words.

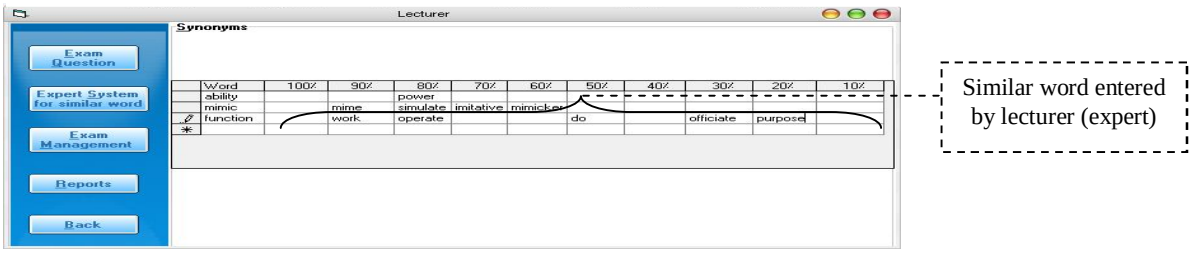


Figure 6. Expert system Interface

IV. RESULTS

The System was developed as a windows-based system that utilizes objects in Visual Basic (VB) for graphical user interface (GUI) construction. MS Access 2002 is selected as the Database Management System (DBMS) of DB system because it is simple and provides handy supports through its help document like many other windows applications, the system uses forms for input and produces output in the form of Reports using special functions in VB. However, the output that does not need to be printed out can be displayed with forms. For its interface design, the system uses several conventions. For example, to prevent too many objects, such as Textboxes, labels, comboBoxes, and others. Frames are used to separate the objects into pages to provide a clearer view as well as avoiding overcrowding. Users can navigate from page to page to carry out their work. Each button is assigned with pieces of code, called Event procedure within form module or User Defined function within Class Module using VB.

The proposed system was developed by integrating several modules as shown in Figure 7.

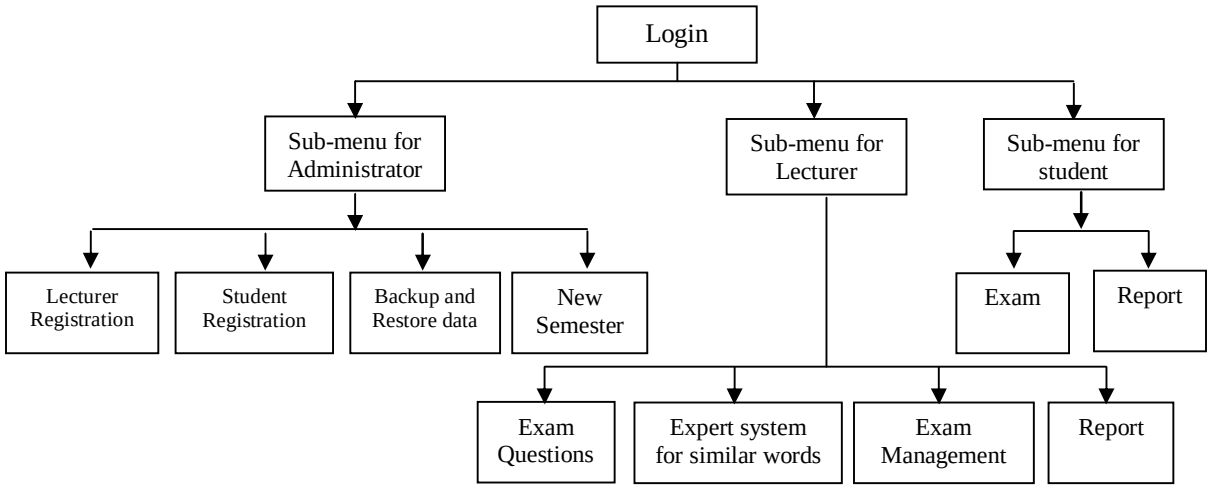


Figure 7. Hierarchy of interfaces within the overall system

A. Exam Questions

As the main objective of the study is to integrate modified BLEU Algorithm with Structured Exam Questions Test Bank and Evaluation System, it is necessary to create exam questions into the DB. To demonstrate how lecturers can add exam questions into question text bank, a snapshot of the interface for adding, modifying and deleting exam question is shown in Figure 8.

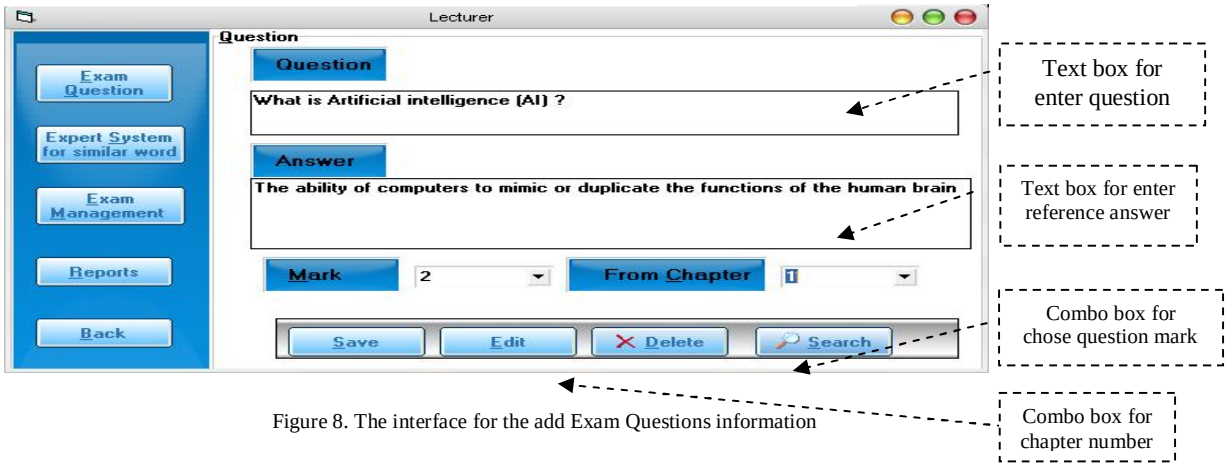


Figure 8. The interface for the add Exam Questions information

To search for the exam questions, the lecturer can click the search button and the table with list of exam question is displayed. The lecturer can then choose the questions from the table; the appropriate information is displayed for further actions. Figure 9 shows the interface for searching Exam Questions information.

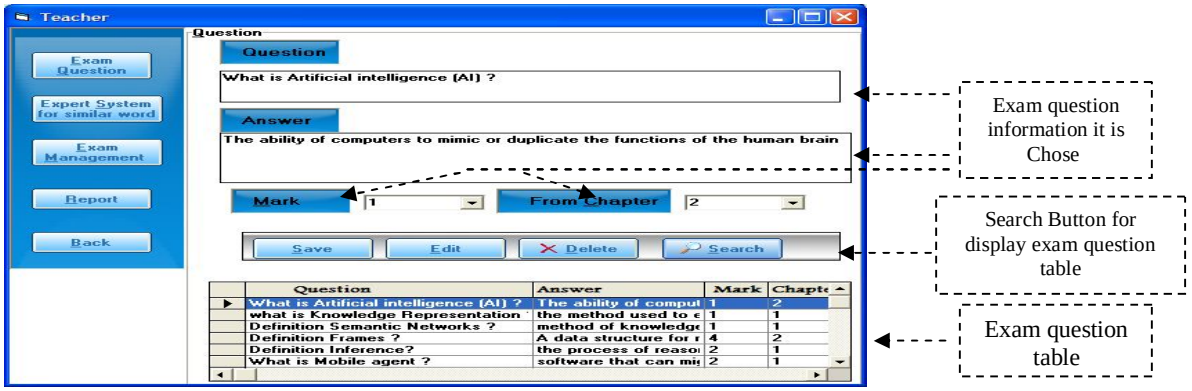


Figure 9. The interface for search Exam Questions information

B. Expert System for Similar Words

One of the main modules in the proposed system *is* expert system that caters for similar words for words that appear in the reference answers. Once the lecturer clicks the save button after filling in the exam question information (see Figure 8), the interface for similar words will be displayed (see Figure 10). The system will only request the user to enter the similar words from the reference answers if the synonyms are not available in the knowledge base of similar words. If the lecturer is requested to fill in similar words table, each cell represents the probability values allocated to similar words ranging from 0.1 to 1.00.

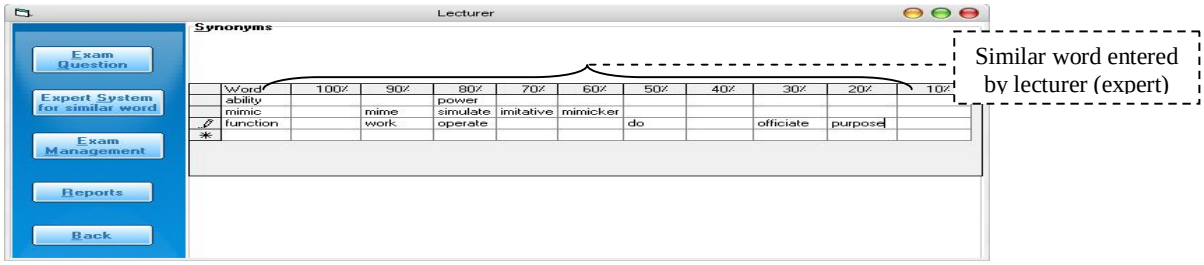


Figure 10. The screen similar word for each keyword enter by lecture

C. Exam management Interface

The exam management window allows the lectures to add, update and delete exam management information. It also provides flexibility in assisting the lecturer to allocate time and appropriate questions from each chapter as shown in Figure 11.

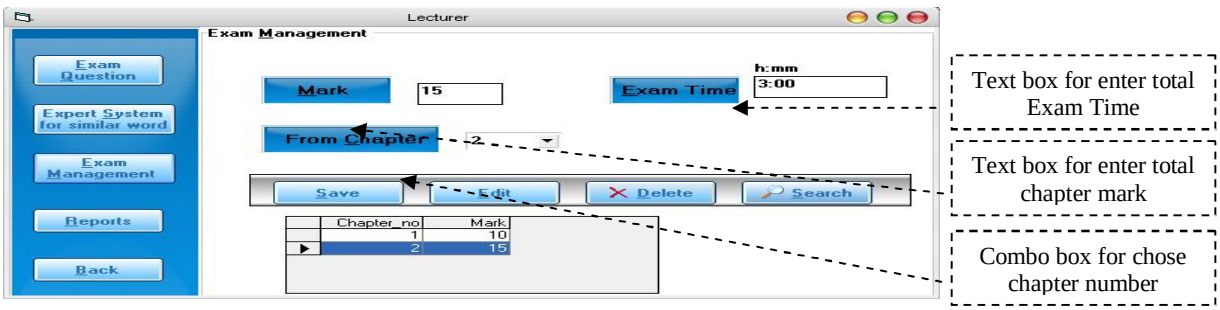


Figure 11. The screen for add new exam management information

D. Student's User Interface (Exam Interface)

In order to use the system, the student must log in by supplying a valid user name and Password. The student's user interface has been designed to show exam questions and view his scores after submitting the examination answers (Figure 12). The modified BLEU module has been implemented to calculate the student's score. However, before the final score is displayed to the student, BLEU module will invoke the expert system in order to check for similar words in the student's answer.

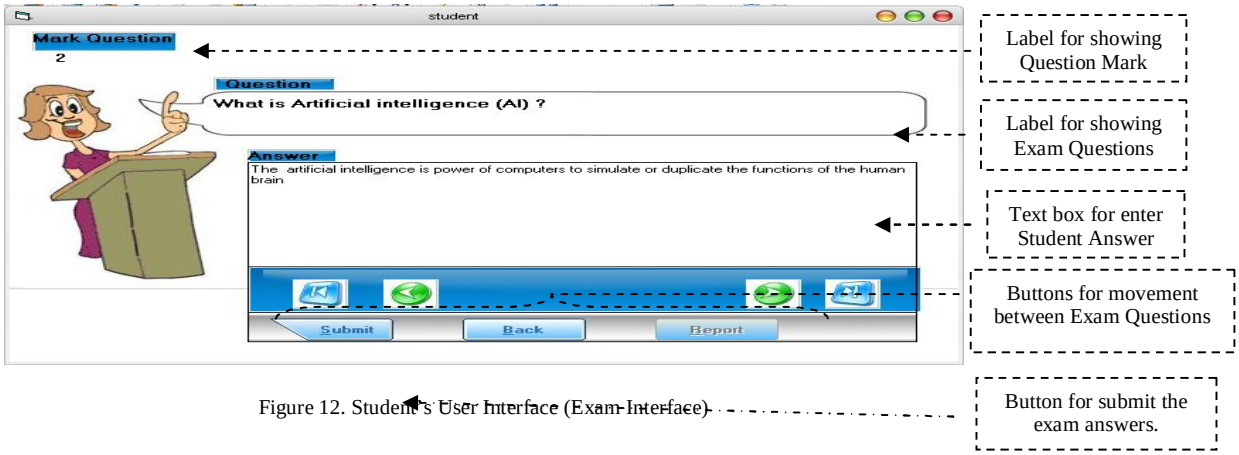


Figure 12. Student's User Interface (Exam Interface)

E. Report

The system results are shown in three different reports, Single Student Report, all student Scores report, student exam answers report. The student results view contains the following information: the points the student has gained for each questions and is shown in Figure 13.

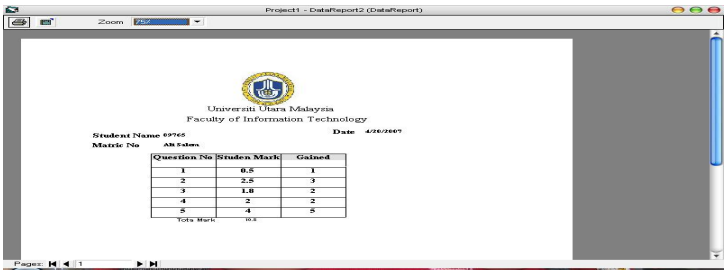
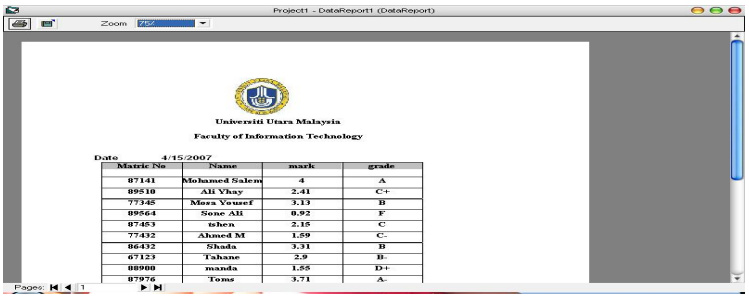


Figure 13. The content for Student Scores report

Figure 14 shows the all students' score in the exam and respective grades.



F. System Testing

The system was further tested with real users (30 lecturers and 200 students). Table 2 exhibits the results of system's evaluation based on some questions on the usage and interface of the system. The overall evaluation of the system indicates that at least 85% of the lecturers and 90% of the students agreed the developed system has shown positive impact on the structured questions marking and writing answers correctly. The advancement in on-learning delivery, automated marking system for non-multiple choices question, the automated marking system for structured and essay type of questions is almost compulsory so that immediate response in marking is vital for the virtual students.

Table -2 System's Evaluation

Users	Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
Lecturers	0%	1%	12%	42%	45%
Students	0%	0%	10%	40%	50%

V. CONCLUSION

The proposed system employs modified BLEU Algorithm and Expert System for similar word to mark open-ended (free-text) responses. It has been designed to facilitate the exam management process, including the students' scores. The lecturer's time can be devoted for other task. In addition, the examination questions test bank and the students can easily be stored for future mining purposes. Since the examination process has been carried out by the system, it is inevitable that such a system can reduce the time for course management, including providing faster feedback to the students. Moreover, the evaluation of this system has been approved by users who have tested this system. The system was found to be able to perform functions correctly as described in the earlier section, flexible and easy to use by the users. Computer-assisted learning can be useful for students, and it is particularly well suited to those which, because of any reason (e.g. being physically impaired) cannot attend traditional lectures.

REFERENCE

[1] D. Whittingdon, H. Hunt, "Approaches to the Computerized Assessment of Free-Text Responses", *Third International Computer Assisted Assessment Conference Loughborough University*, June 1999.

[2] D. Perez, E. Alfonseca and P. Rodriguez, "Application of The Bleu Method for Evaluating Free-Text Answers in an E-Learning Environment", *to appear in Proceedings of LREC-2004*, Lisbon, 2004.

[3] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation", *Technical Report RC22176 (W0109- 022)*, IBM Research Division, Thomas J. Watson Research Center. 2001.

[4] T. Szymanski, "Recognizing Textual Entailment with a Modified BLEU Algorithm", *Available from: https://www.researchgate.net/publication/266017019*

[5] E. B. Page, "The Imminence of Grading Essays By Computer", *Phi Delta Kappan*, 1966.

[6] P. W. Foltz, D. Laham, and T. K Landauer, "Automated Essay Scoring: Applications to Educational Technology". *Proceedings of ED-MEDIA '99 Conference, AACE, Charlottesville*, USA. 1999

[7] Vantage Learning Tech, "A Study of Expert Scoring and Intellimetric Scoring Accuracy for Dimensional Scoring of Grade 11 Student Writing Responses" *(Report no. RB-397)*. Newtown, PA: Vantage Learning. 2000

[8] J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Bradenharder, and M. D. Harris, "Automated Scoring Using A Hybrid Feature Identification Technique", *in Proceedings of the Annual Meeting of the Association of Computational Linguistics*, 1998.

[9] J. Burstein, C. Leacock, and R. Swartz, "Automated Evaluation of Essays and Short Answers", *in Proceedings of the International CAA Conference*, 2001.

[10] J. Burstein, M. Chodorow, C. Leacock, "Criterion^{SM} Online Essay Evaluation: An Application for Automated Evaluation of Student Essays", *in Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, Acapulco, Mexico, August, 2003.

[11] L. S. Larkey, "Automatic Essay Grading Using Text Categorization Techniques", *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 90–95). DOI: 10.1145/290941.290965. 1998

[12] A. Mikhailov, "Indextron: Intelligent Engineering Systems through Artificial Neural Networks" 8, 57. 1998

[13] J. R. Christie, "Automated Essay Marking - For Both Style and Content". 1999.

[14] P. Y. Ming, A. A. Mikhailov, and T. L. Kuan, "Intelligent Essay Marking System". *in C. Cheers (Ed.), Learners Together, Feb. 2000, NgeeANN Polytechnic, Singapore. 2000.*

[15] P. Dessus, B. Lemaire and A. Vernier, "Free Text Assessment in A Virtual Campus", *in Proceedings of the 3rd International Conference on Human System Learning, pp. 61–75. 2000.*

[16] D. Callear, J. Jerrams-Smith and V. Soh. "CAA of Short Non-MCQ Answers". *in Proceedings of the 5th International CAA conference, Loughborough, UK. 2001.*

[17] T. Mitchell, T. Russell, P. Broomhead, and N. Aldridge, "Towards Robust Computerized Marking of Free-Text Responses", *in Proceedings of the 6th CAA Conference, pages 233–249, 2002.*
[18] L. M. Rudner, and T. Liang, "Automated Essay Scoring Using Bayes' Theorem". *The Journal of Technology, Learning and Assessment, 1(2), 3-21. 2002.*
[19] R. M. Carro, E. Pulido, and P. Rodríguez, "Dynamic Generation of Adaptive Internet-Based Courses", *Journal of Network and Computer Applications*, 22 (4), 249-257. 1999.
[20] A. Ortigosa, R. M. Carro, "The Continuous Empirical Evaluation Approach: Evaluating Adaptive Web-Based Courses", *International Conference on User Modeling, Springer Berlin Heidelberg*, pp. 163-167. 2003.
[21] O. Mason, and I. Grove-Stephenson, "Automated Free Text Marking with Paperless School", *in M. Danson (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough University, Loughborough, UK. 2002.*
[22] B. Bloom, "Taxonomy of Educational Objectives: The Classification of Educational Goals", *Handbook I, Cognitive Domain*. New York; Toronto: Longmans, Green. 1956.
[23] J. Z. Sukkarieh, S. G. Pulman, and N. Raikes, "Auto-Marking: Using Computational Linguistics to Score Short, Free Text Responses", *Paper presented at the 29th Annual Conference of The International Association for Educational Assessment (IAEA)*, Manchester, UK. 2003.
[24] Ros´e, C., Roque, A., D., D. B. & VanLehn, K. (2003), 'A hybrid text classification approach for analysis of student essays', Build Educational Applications Using Natural Language Processing pp. 68–75.
[25] S. Dikli, S. Bleyle, "Automated Essay Scoring feedback for second language writers: How does it compare to instructor feedback?"*Assessing Writing 22*, pp. 1–17. 2014.
[26] J. Wilson and A. Czik, "Automated Essay Evaluation Software in English Language Arts Classrooms: Effects on Teacher Feedback, Student Motivation, and Writing Quality", *Computers & Education*, 100, 94-109. 2016.
[27] M. Warschauer, and D. Grimes, "Automated Writing Assessment in The Classroom", *Pedagogies: An International Journal, 3, 22e36.* 2008.
[28] D. Grimes, and M. Warschauer, "Utility in A Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation", *Journal of Technology, Learning, and Assessment, 8(6), 1e44. Retrieved December 12, 2014, from http://www.jtla.org*
[29] F. Kiyoumarsi, "Evaluation of Automatic Text Summarization Based on Human Summaries", *2nd Global Conference on Linguistics and Foreign Language Teaching, LINELT-2014*, Dubai, UAE, 2014.
[30] H. Moena, L. Maria, D. Peltonen, J. Heimonen, A. Airola, T. Pahikkala, T. Salakoski, S. Salanterac, "Comparison of Automatic Summarisation Methods for Clinical Free Text Notes, *Artificial Intelligence in Medicine 67*, pp. 25–37, 2016.
[31] M. Hermet, S. Szpakowicz , L. Duquette, "Automated Analysis of Students' Free-text Answers for Computer-Assisted Assessment", *TALN 2006*, Leuven, pp. 835-845, 2006.
[32] J. Nunamaker, M. Chen, and T. Purdin, "System Development in Information Systems Research", *Journal of Management Information Systems, 7:3*, pp. 89 – 106. 1991.