A SURVEY OF COMPUTER-BASED ESSAY MARKING (CBEM) SYSTEMS

Sa’adiyah Darus
Siti Hamin Stapa
Supyan Hussin
Koo Yew Lie
Faculty Of Language Studies
Universiti Kebangsaan Malaysia

Abstract
The following survey is presented by firstly looking at the initial research carried out in the area of computer-based essay marking systems. Early researchers, inspired by natural language processing, studied how the computer might play a role in evaluating students’ writing. Computer-based essay marking systems developed from late 1960’s, attempted to prove that a computer can mark essays as good as any other human beings. These are Project Essay Grade-1, 2 and 3. As time goes by, more sophisticated systems are developed, such as A Simple Text Automatic Marking System, Intelligent Essay Assessor, e-rater and the program developed by S. L. Larkey. Some of these systems have explored the use of advanced computational linguistics and statistical techniques for automatically scoring a variety of writing responses. Most of these systems are promising at the research level. The paper conclusively indicates that further research needs to be carried out in order to make these programs commercially available to instructors.

Introduction
The computer was initially developed for handling numerical calculations. However, as time goes by, and as the computer gets cheaper and cheaper, other computer applications are being developed.

One of the computer applications that are widely being used is in the area of document preparation. A number of document preparation support tools are available in the market. Some of these tools are only providing support for document formatting, while others provide support authors in writing. We call them as author support tools (AST). The first effort in providing these AST was perhaps, developed within the UNIX operating system. The UNIX word processing system is known as NROFF and is supported by some other tools, for example Style and Diction. With the advent of microcomputers, new word processors have been developed. One of the
most popular word processors was WordStar. Others are Microsoft Word tools such as “Revisions” and “Track Changes”.

Apart from the tools that support authors in writing and revising, activities that define process-based pedagogy. There is a need to develop tools to help document assessors in evaluating the quality of a document. This is particularly important, especially for instructors in evaluation of students’ essays.

Essay writing is an important aspect of learning, and it is especially so at the tertiary level of education. It is usually given in the form of writing assignment and it comprises of a major component of any course. Students are required to submit their writing assignments by a given deadline. Instructors would then need to handle and mark these assignments.

Handling writing assignments is a difficult and tedious task for instructors especially when they have to mark long essays in vast numbers. With the traditional way of evaluating essays, instructors are faced with two major problems. Firstly, the management of essays. Instructors need to keep track of assignments that are not submitted, late submissions and missing scripts. In addition, marks need to be tabulated. Secondly, marking of scripts. All assignments submitted must be marked by a certain deadline. Marking has to be done as soon as possible, so that prompt feedback can be given to students. At the same time, instructors need to be consistent in marking, so that grades given are fair and accurate.

With the increased availability of computer technology, it is important for us to look into the possibility of using computers in handling and marking essays. The objective of this paper is to document work done by previous researchers in the area of computer-based essay marking. It is hoped that it will be of benefit to future researchers to undertake further work in this area. Some tools that have been developed can be divided into two categories: semi-automated and automated system (Sa’adiyah, 1999).
Semi-Automated Systems

In a semi-automated system, the computers carry out only some aspects of the handling and/or marking. Instructors are needed to complete the rest of the process. This is exemplified in the following systems.

Methodical Assessment Of Reports By Computer

Methodical Assessment of Reports by Computer (MARC) is based on a program known as REPORT that was created by Marshall (1985). REPORT is applicable to a large variety of reports, while MARC is confined to marking the “glass factory” reports (Marshall and Barron, 1987). Both programs were developed at the Department of Language and Social Science, Papua New Guinea University of Technology.

MARC is a report-marking program for engineering course at the university, that enables the instructors to provide three pages of individualised feedback on every student’s report, based on a certain criteria for report writing. The program comprises of three main parts: a database, an evaluator and a language processor. By using MARC, the students can be assured of the consistency of marking.

Before marking begins, the computer is provided with some general information, for example the number of students, the name of the class and the marking scheme to be used. Then, the student’s name is obtained and printed on the feedback sheet. From this point onwards, the computer asks questions about each section of the report, and then prints the appropriate feedback given by the instructors. After completing all the questions relating to the “main body” of the report, it continues to another section, the “coherence” of the report and the “diagrams”. It then continues to ask questions about grammar, length of sentences and spelling. Finally, the computer suggests a mark for all these features, and prints the appropriate feedback and the overall mark.

Markin 32

Markin 32, which is developed by Holmes (1996), is a correction tool that allows instructors to mark written work submitted by students in the form of...
electronic documents. Markin 32 Version 1.2 provides five marking facilities; annotation buttons, add feedback, add comment, add a grade, and compile error statistics. It allows instructors to return marked written work to students in three different formats: Web page (HTML document), word-processor file (Rich Text Format document), and text file. Students are able to access the feedback given by instructors on the Internet.

The corresponding author has encouraged her students to use this particular software in her Written Communication classroom at Universiti Kebangsaan Malaysia. Markin 32 helps the students in completing their task of writing journals. By using this tool, 87.5% of the students are interested to read the feedback of the writing that they had submitted and 50% of them found the format of the feedback, in this case Web page format, had been of great help. At the end of the course, 62.5% of the students totally agreed that they had more confidence to write using this tool.

Markin 32 Version 2.0 has just been released and is available through the Internet (Holmes, 2000). It incorporates many changes and enhancements. Some of these new features include unlimited annotation buttons (negative and positive buttons) and free floating button bar. Dragging and dropping can move these annotation buttons, while annotation key can be automatically generated. Best of all, it is compatible with Windows 2000 and Office 2000.

**Automated Systems**

Automated systems use computers in handling and marking essays automatically. Some of these systems are as follows: Project Essay Grade, A Simple Text Automatic Marking System, Intelligent Essay Assessor, e-rater and the system proposed by S. L. Larkey.

**Project Essay Grade**

Project Essay Grade was developed by Page et al (1968). They conducted the first study in the area of automated essay marking, by defining a large number of objectively measurable features in the essays, such as essay length, average word length, average sentence length, proportion of prepositions and common words etc.
and used a standard **multiple linear regression** to try to predict the scores that human graders would give to these essays. Project Essay Grade (PEG)-1 was developed at the University of Connecticut, U.S.A. The research was carried out on a set of student essays that had already been graded manually by English teachers. The correlations of the computer program with four human judges were 0.50 (Page 1994).

A subsequent study, PEG-2 used **multi-trait ratings**. The number of judges in this study was doubled and it was shown that the computer was more accurate in scoring creativity and style than in scoring mechanics when compared to human judges.

The third version of PEG explored **subject-matter essays**; science, social studies, English and foreign language (Ajay, Tillet, and Page, 1973; Page, 1985; Page, Tillet, and Ajay, 1989). The U.S. Office of Education supported the research. PEG-3 uses the same sort of regression design as in the previous projects with additional features: specific content words and their synonyms. The research was successful and it involved only one human grader for each of the subject-matter essays.

Page and Peterson (1995) positively claimed that PEG program could become a cost-effective alternative to human ratings of essay tests. It was in this research that they established the fact that the system would work in practice where they made use of 1,314 essays composed on computers by college students.

Recently, Hiller (1998) has proven that PEG can evaluate the **quality of writing** products namely, instructional texts or regulations. It can also be used to monitor various aspects of style, for example emphasising short sentences, using familiar and concrete words, being specific, avoiding exaggeration and avoiding vagueness. It is important to note that through out the researches, the PEG’s operating assumption has been that appropriate human judges must define the true quality of essays.

**A Simple Text Automatic Marking System**

Research in the area of natural language semantic understanding at the University of Nottingham, United Kingdom has made A Simple Text Automatic Marking System (STAMS) available for researchers (Lou and Foxley, 1994).
The model adapts the online Roget’s Thesaurus database to get the meaning of words, uses fuzzy logic techniques to produce fuzzy semantic sentence pattern, measures the semantic distance to decide the closeness of two sentences, and imitate human behaviour in the language processing model, in order to build a loose grammar rule and decreases the ambiguity of the meaning of the sentences (Lou, 1996).

**Intelligent Essay Assessor**

On the contrary, the research team at the University of Colorado, headed by Professor Landauer, took a different approach in CBEM. They came up with Intelligent Essay Assessor (IEA), which uses “Latent Semantic Analysis” (LSA). LSA is a fully automatic mathematical or statistical technique for extracting and inferring relations of expected contextual usage of words in passage of discourse.

In many aspects, the program is different from the one developed by Page (1994) and Lou (1996), in that it does not involve natural language processing or artificial intelligence; it uses no humanly constructed dictionaries, grammars, syntactic parsers, or the like. It takes as its input only raw text parsed into words defined as unique character strings and separated into meaningful passages such as sentences or paragraphs (Landauer, Foltz, and Laham, 1998). LSA uses a k-nearest neighbour algorithm to access the k essays most similar to the new essay to be graded. It measures the knowledge content of essays and is not designed to grade stylistic considerations like grammar and spelling.

In order to utilise the software, one has to “feed” it with information about a topic. Then, it learns from the text and assigns a mathematical degree of similarity or “distance” between the meaning of each word and any other word. The instructors can evaluate essays in two primary ways. Firstly, a lecturer needs to grade enough essays to provide a good statistical sample and then use the software to grade the remainder of the essays. The second method compares all the student essays to a single expert’s essay. It can then tell students what important subject matter was missing from their essays. They argue that IEA enables essay grading in large scale testing for example, introductory college classes or standardised testing for entrance to professional schools (Murray, 1998).
E-rater

Other organisations that are interested in using the computer to mark students’ essays are the Educational Testing Service (ETS), Princeton NJ and Hunter College, New York. They have developed a system known as Electronic Essay Rater (e-rater) based on writing characteristics specified in the holistic scoring of Graduate Management Admission Test (GMAT) essays (see the GMAT Web site at http://www.gmat.org).

E-rater makes use of advanced computational linguistics and statistical analysis techniques for automatically scoring a variety of constructed writing responses (Burstein, Kukich, Wolff, Lu and Chodorow, 1998a). It is able to analyse essay features such as discourse, syntax and topical content (Burstein, Kukich, Wolff, Lu, and Chodorow, 1998b).

It makes use of a lexicon based on the conceptual framework of conjunctive relations from Quirk et al (1985) to recognise discourse cues. In order to recognise syntactic structure, e-rater uses a parser that is able to identify subjunctive auxiliary verbs (e.g. would, should, might) and complex clausal structures (e.g. complement, infinitive). In order to capture use of vocabulary or identification of topic, it uses content vector analyses that are based on the vector-space model commonly found in information retrieval applications. Thus, it is claimed that the system can be used to evaluate non-native speakers’ writing, as these features are not confounded by non-standard English syntactic structures or stylistic discourse structures (Burstein and Chodorow, 1999).

Program Developed By S. L. Larkey

The other institution doing research in CBEM is the Centre for Intelligent Information Retrieval, University of Massachusetts. Larkey (1998) uses an approach whereby binary classifiers are trained to distinguish “good” from “bad” essays. The scores output by these classifiers are then used to rank essays and assign grades to them.

Correlations in this study are generally in the high 0.70’s and 0.80’s, depending upon essay type and upon the quality of the human ratings. These levels are
comparable to those attained by Landauer et al (1997) and Page (1994). These correlations seem high, and are comparable to the correlations between human judges.

**Conclusion And Discussion**

Computers have come a long way since it was first invented. It has gone through tremendous development especially in the area of computer-based essay marking for the last forty years or so. With the advent of Information and Communication Technology, it will continue to do so in future.

The literature review clearly indicates that semi-automated and automated systems in handling and marking students’ writing are available and seem promising at least at the research level. However, for some reason or other, none of these programs are commercially available, except for the Intelligent Essay Assessor, developed and marketed by the Knowledge Analysis Technologies, LLC Boulder, Colorado, USA (see IEA website at [http://www.knowledge-technologies.com/](http://www.knowledge-technologies.com/)). As such further research in this area need to be undertaken to resolve why this is so.

Possible areas that researchers need to look at are: theory and practise of writing; issues in assessment of writing for example, politics, policies and practices; composition pedagogies; genre analysis and cross cultural rhetoric. As for utilising CBEM system for the Malaysian environment, researchers specifically need to look into the writing needs of our students from various racial and ethnic backgrounds and also the writing requirements for Malaysian undergraduate degrees.

In order to ensure that instructors can continue to evaluate students’ essays effectively in the new millennium, it is evident that future research in computer-based essay marking, that is customised for our Malaysian instructors could eventually prove useful for education in Malaysia.

**References**


