Design and Integration of an Automated Assessment Laboratory Experiences and Guide

Heinz Dreher and Naomi Dreher
Curtin Business School
Heinz.Dreher@cbs.curtin.edu.au
Naomi.Dreher@cbs.curtin.edu.au

Torsten Reiners
University of Hamburg
reiners@econ.uni-hamburg.de

Abstract: The aim of the Automated Assessment Laboratory (AAL) being established at the Curtin Business School (CBS) is to provide lecturers with the opportunity to have essays automatically assessed using MarkIT. This automated essay grading tool is most suited to those units that have large numbers of students. In this contribution, we describe our approach to design and integrate the AAL in the curriculum, report our experiences and provide a guide for other institutions.

1. Introduction

The purpose of the Automated Assessment Laboratory (AAL) is to assist academic staff of CBS running units with large enrolments to consider and/or trial the use of Automated Essay Grading (AEG) technology. The aim of the AAL is to provide lecturers with the opportunity to have essays automatically assessed using MarkIT developed by Robert Williams and Heinz Dreher from the School of Information Systems at Curtin University of Technology. MarkIT provides a numerical essay score and comprehensive visual and textual formative feedback on essay content, which enables the lecturer to discuss with each student the strengths and weaknesses of their essay and suggest areas of improvement. This diagnostic use of assessment to provide feedback to students over the course of instruction is useful formative assessment. MarkIT provides the opportunity for immediate evidence of student learning and development (Dreher 2006). It also enables lecturers to examine if the learning goals and objectives are met in all sections of the assessment task.

Nowadays most systems in the field of e-learning rely on evaluation and assessment by posing questions in computable formats like multi-choice or well-defined answers. But even clozes can cause false interpretation if even simple tasks as the check for synonyms are not performed. On the other hand, essays are required to verify knowledge and its application to new problems (transfer of learning) as well the capability to express (new) arguments. In addition, the student is more challenged when answering essay-type questions and therefore more motivated rather than repeating numerous right and wrong answers. From the perspective of the evaluator, multiple-choice exams have advantages like simple and fast to grade and non-negotiateable for individual cases. On the other hand, they are far more difficult to create than questions requiring an essay as an answer. This disadvantage in essay answers is that there is not the “one” answer that is right and it is far more challenging for the evaluator to be objective, especially if two answers are completely different but both are valid solutions. Regarding the large variety of other influences like style of language, form and handwriting, clarity of expression, errors in grammar, nationality and cultural background, or even having an order in which the essays are marked (fatigue of the marker) effects the result due to human marker. There are many influences causing subjectivity in grading. As shown by (Williams and Dreher 2004, Williams 2006), human graders who grade the same exams might result in a correlation only between 0.78 and 0.81. In comparison, the correlation between the human and computer grade is 0.79 and, therefore, competitive; see Figure 1 for a visualization of the results of one essay grading trial. Figure 2 shows another example where the human and computer grader have a correlation of 0.72.

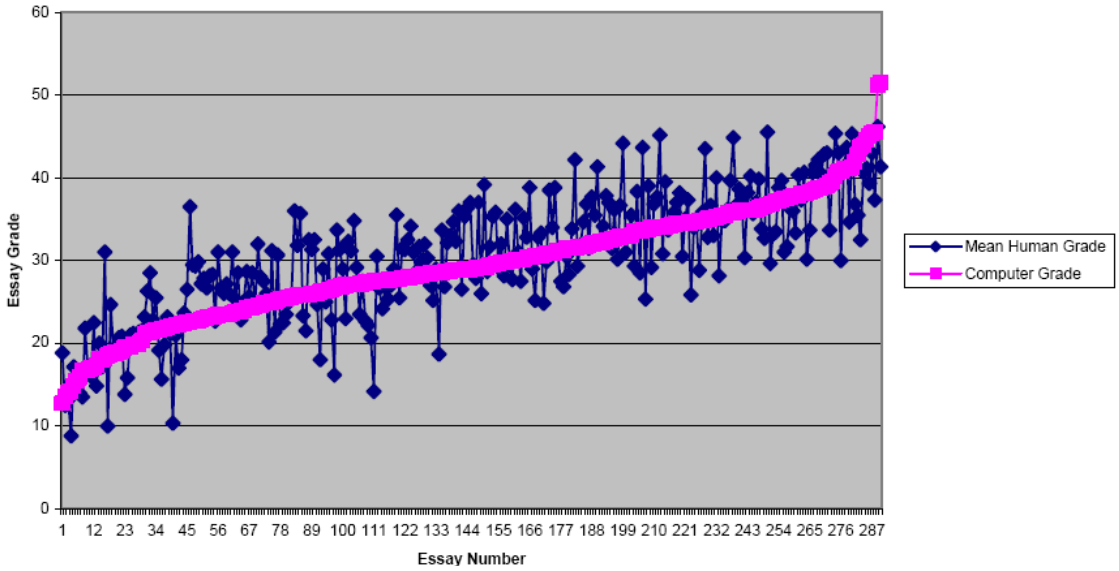

Figure 1: Grading of 290 Essays with three human graders versus the computer grade. Source: Williams, 2006

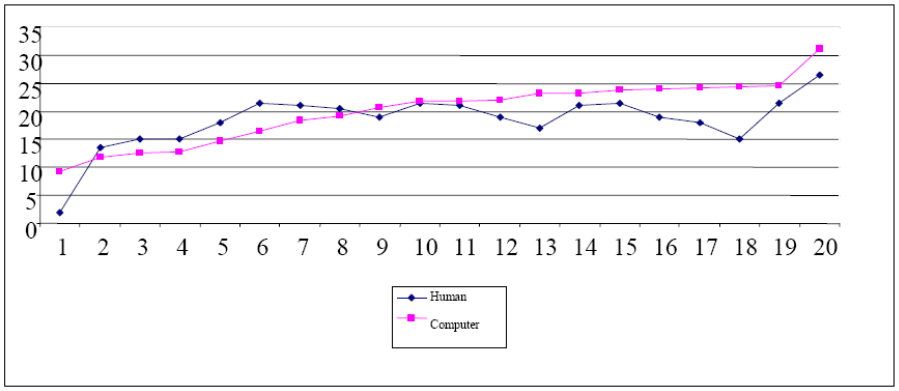

Figure 2: Results of grading business law essays by human and computer. Source: Williams & Dreher, 2004

After discussing the human-grader, let's go back to the computer as an alternative. The computer gets the essay in electronic form so that optical factors (handwriting, style, keeping a certain form) as well as the background of the student should not have an influence. Nevertheless, the quality of the essay has to be quantified using different metrics, which will be influenced by the capability of the student. Note, that the final grade is composed by several factors so that metrics evaluating the language and their usage might have only a small impact. The evaluation of the answer compared to a model solution is the subject of the AEG algorithm, where the meaning is analyzed and compared – model against student essay. This is approach does not involve subjective decision, but one has to verify that the computer interprets the answer correctly as a valid solution.

A model answer is prepared by the instructor that contains the core knowledge required to achieve a 100% (or the high score). The system may eventually be able to grade a student essay against a number of model answers, in case the instructor wishes to use numerous content models. The instructor can also provide about 100 - 200 human graded essays (ideally each graded by three humans) and their scores, for training purposes. Experience has shown higher computer-human correlations are possible with the use of the training essays. These model and training answers are processed as described above. The system then performs a content matching task in which the model answer content summary is compared against each of the training essay content summaries. Many aspects of the relationships between the model and training essays are then computed, and a linear regression model computed to derive a scoring equation. Unmarked student essays are then processed to build the content summaries that are to populate the essay content matrix according to the dimensions revealed in the linear regression phase. Finally, the scoring equation is used to produce a score for each essay.

The following section discusses the components required to realize and operate an Automated Assessment Laboratory (AAL). In Section 3, we discuss the different roles and their tasks in the AAL, before we continue with a short discussion about benefits and problems in Section 4. We conclude the paper with a short Outlook with the first field test being described.

2. Components of an AAL

The AAL is about providing a platform for learners as well instructors to minimize the amount of work and maximize the success of the assessment. Therefore, general guidelines are required to support the whole laboratory and guarantee a high quality level in grading and archiving/recording for later reviews. In addition, we discuss our guidelines in this Section; see also (Wulf 2004, Almond et al. 2002) for further options. Note, that we consider especially essay assignments and, therefore, assume that the submission consist only of one document. In general, it is important to bundle all files belonging to one submission, e.g. in zip-archives.

- (Document/Content) Management System, automatic document analysis and transformation

Even though the structure of documents for the submission is announced and clearly described within the classroom, students do not necessarily follow the guidelines in all aspects. Therefore, an analysis needs to be performed, which will – depending on the outcome – be used for later transformations. To increase the performance and quality of the AEG, we assume the following properties for a submission:

- Simple document without fancy formatting rules. For the transformation, the text must be extracted in a ASCII-format where only headers, paragraphs and enumerations are given
- Tables and Figures should not be included as they cannot (yet) be interpreted or compared
- No equations

The information about students who submitted the document has to be included and stored. On the one hand, the information can be encoded in the file name of the document (e.g. Lastname_Firstname_ID_ Course_Assignment.format) or in the document using a predefined structure. The disadvantage is that in case of errors, it requires much more work to perform corrections. On the other hand, a special submission system could be used, either based on a web-site or integrated in a virtual learning environment. Here, the students generally have done a previous registration so that their data is already given and can automatically be included in the submission.

- Database to store assignments/grades of several courses
- AEG-Software (MarkIT)
- Workflow/Task -Management (Assignment of exams to grader)

Storage in a hierarchical file system vs. database: The advantage of the first approach is the simplicity in starting and extending an archive without requiring any further knowledge or software. The folder represents the levels in the hierarchy. The top level can be used for the semester, below are the course name, instructor, and assignment. The lowest folder contains the documents. Another approach would include a database, where the essays are either stored or referenced by keeping the location in the file system.

To keep several versions of submitted documents, the filename needs to be extended by a version number. Using a hierarchical file system instead of database will result in large quantity of files and, therefore, complexity. Resubmission might be necessary in case of errors by the students before the deadline or by submitting a corrected version after a first grading

Storage of the results can either be done in a database or encoded in the file name in form of a number representing the percentage (as it is currently done). The latter approach has disadvantage as handling of the added information is difficult to use in cases like curving. Otherwise, the grade can be stored together with further information about the grading in a database Remarks are stored in an extra document assigned to the file by matching file names.

Further tools are:

- Interface to access the documents in a general form, i.e. for integration in other application or web-interfaces.
- Organization of students, assignments and results
- Tools to interpret the results and to comment the submitted essays
- Tools to perform the grading and to edit the concepts of MarkIT.

- Handling the experiments, especially the training sets

3. Roles required for operating the AAL

With respect to the processes within the AAL, we have to distinguish the following roles being involved for unobstructed operations. Note that all roles require further staff for support but will not explicitly be discussed in this section.

- **Head of Lab**: Management and validation of the progress, daily business, and financial concerns. The budget has to be procured and distributed to all cost positions. Therefore, new proposals to finance the daily business and the ongoing research for improvements and upgrades have to be prepared and submitted to the university (executive dean) or for other grants. With respect to the services offered by the lab, the head has to establish the selling price in accordance to the lecturer, type and size of exam and the actually performed services.
  The lab has to be financed and requires customers, i.e. instructors and their students. Thus, material describing the lab in form of promotional pamphlets and posters needs to be designed and distributed within the institutes of interest. In cases where all process can be done electronically, the material can also posted via e-mail and web-sites. In case of interest, meetings with the instructors have to be organized and arranged.

- **Education Consultant**: Mainly supporting the head of AAL in respect to achieve background knowledge about the offered grading services and advising instructors to use the AAL for the grading process. Especially questions about the applicability and reliability have to be answered and demonstrated by historical experiments and test results.

- **Lab Administrator**: The lab administrator takes responsibility for the whole lab, its organization as well as equipment.

- **Technical Administrator**: Taking care of all tasks being connected to hardware and software.

- **Developer**: The AAL needs to be updated, upgraded, and improved by gradual or innovative extensions. Here, the newest products and (scientific) literature has to be reviewed and considered for integration in the AAL. Besides general subjects, there should be a focus on innovative themes like automated essay grading to improve the grading by a large extend.

- **Instructor**: The instructor is preparing the students for the exam, homework and self-tests. Therefore, this role will have a great influence on the outcome. Within the AAL, the main role is about providing the control model.

- **Grader**: The grader must provide a true foundation for the first X exams to which the other are later related to. That is, the grader needs to be objective and constant over time with respect of fatigue and passing different grades to a comparable work.

- **Student**: Basically the one who writes the text to be graded. Knowing about the kind of grading might influence the style of writing, i.e. to influence the final outcome.

4. Benefits and Problems for Lab Personal, Lecturers and Students

Preliminary informal conversations with academics and lecturers as well as feedback from instructors showed reluctance to use AEG within the AAL. Main arguments are:

- Skepticism related to academic benefit and integrity
- The lack of useful, conceptual, written feedback
- Inappropriate current assessment items that would not be suitable for AEG
- The inappropriateness of such a tool at the university level

Even though the methods showed very good results in past experiments, the doubts are understandable and have to be considered. In education – independent from the kind and level – the automation of the grading process is a

sensible field as most participants belief that only exams using, e.g., multiple-choice can be graded by computers as they do not provide features as intelligence and understanding of written answers. A human grader might see the solution hidden in weird formulations. We did not intend to persuade instructors but rather invited them to have a look at our results so far and to perform the evaluation in parallel to human graders. Afterwards, they can decide based on the outcome if they continue to corporation.

Nevertheless, there are two important issues to bear in mind: (1) the exam needs to be structured in a way that it can be used with the Software MarkIT and (2) the feedback various – for each exam – from the individual remarks that are generally done by the human grader. It is possible to argument the grade and to visualize the distribution of points. Thus providing profound feedback and suggestions to generally improve an assignment is not yet given. As we described before, the grading is based on model answers and, therefore, the freedom has to be restricted. For example, it would be difficult to computer-grade essays where the student can select a subject related to the class by him or can choose a publication

Another problem for the educational consultant was to have the instructors to do their share of work as they complained about a generally high workload and participation is not possible as it would require further time, no urgency as the exams were also human graded and, therefore, being returned to the students and the trial status took some seriousness from the involvement.

Some of these problems result from the experimental stage. As soon as the AAL is established and productive, there are several benefits – and definitely further problems – for all roles:

- The AAL allows consistent and interactive formative assessment with immediate feedback for the students instead of waiting several days or weeks for human grades. In addition, the AAL can be used in self-tests in e-learning scenarios, which is otherwise only possible if multiple-choice or other simple question types are used (strategic and competitive advantage)
- Improving the quality of teaching, i.e. consistent assessments and comparable results due to application of strict rules equal for all students (monitoring).
- Reduction of costs as the AAL charges less per exam compared to human graders
- Electronic handling of exams and, therefore, less paper
- Prejudgment is eliminated or lowered, respectively, during grading
- Detection of plagiarism is simpler as assessments are electronically submitted and the MarkIT-grading is planned to be used in the future with respect to the methodology

It should be mentioned that even a support for the human grader by the AAL can be useful as the advantages of the electronic submission can be used for the human grading by, e.g., presenting all questions of one type.

5. Conclusions and Outlook

The following table shows subjects and number of participating students currently involved in preliminary experiments. As the AAL is currently in the formation phase, the results are used to get further experience with real exam settings. Note, that the final submission will include results from the ongoing semester.

Course	# Students	# human graded	# AEG
Business (BIS 100)	~600	~600	~600
Accounting (ACC100)	~400	200 + 200	200 + 200

For ACC, the student's assessment was based on a use case and had two options: either (1) finding two weaknesses and recommending a solution using 500 words or (2) find two strengths that can be used to improve the profits. The instructor first decided to have 200 assessments human-graded and computer-graded and based on the result, have the second half either graded with both methods or only by the computer. With respect to verify the quality of computer-grading, all 400 essays for also human graded. For BIS and Small Business, it was decided to grade all essays by human and computer. In BIS, the students had two possible subjects to choose from. The results are presented within the presentation.

The paper describes an AAL using AEG to improve the assessment of students. We presented an overview of the required components, the roles to have staff for, and the benefits and problems, which can occur. In addition, we describe how the multilingual and multi-cultural backgrounds can be considered and the software extended

6. References

Almond, R.; Steinberg, L. & Mislevy, R. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. Journal of Technology, Learning, and Assessment, 1, 4-63.

A.W.Chickering, Z.F.Gamson (1987). Seven principles for good practice in undergraduate education. *American Association of Higher Education Bulletin*, 3-7.

Dreher, Heinz. "Interactive on-line formative evaluation of student assignments." *Journal of Issues in Informing Science and Information Technology* 3 (2006): 189-197.

R.Williams and H.Dreher (2005). Formative assessment visual feedback in computer graded essays. *Journal of Issues in Informing Science and Information Technology* 2. 23-32.

Wulf, T. (2004). Using learning management systems to teach paperless courses: best practices for creating accreditation review record archives. Journal of Computing Sciences in Colleges, Consortium for Computing Sciences in Colleges, 20, 19-25.

R.Williams (2006). The power of normalised word vectors for automatically grading essays. *Journal of Issues in Informing Science and Information Technology* 3. 721-728.

R.Williams, H. Dreher (2004). Automatically Grading Essays with Markit? *Journal of Issues in Informing Science and Information Technology* 1. 0693-0700.