# Text Mining for Student Assessment

Robert W. Reimer, *UGRU IT Services, E-mail: bob.reimer@uaeu.ac.ae*
*U.A.E. University, Al-Ain, P.O. Box: 17172, U.A.E.*

## Abstract

A portion of the University General Requirements Unit's program is dedicated to instruction in English writing. Student assessment in the writing program generates a large amount of marking for instructors. A number of research and commercial packages have been identified that could assist in automated grading, but until recently the expense and effort required to use them seemed too great.

A number of methodologies have been used by previous researchers in automated grading of essays. The literature does not appear to include any references to the use of Classification Association Rules Mining (CARM) to the problem of automated essay grading.

This project extends UGRU's existing proprietary suite of computer-based assessment tools to collecting and grading of essays. It then investigates the applicability of CARM techniques to assigning a coarse grade to the essays using commercially available data mining tools. Mining models based on strictly the text submitted by students, on strictly linguistic attributes of the text, and on a combination of the two sets of attributes were examined. Evaluation of the mining classification shows that it is possible that CARM techniques could be usefully applied to the problem of automated grading of student essays.

## 1. INTRODUCTION

The University General Requirements Unit (UGRU) is responsible for providing the first year preparatory program for United Arab Emirates University (UAEU). In order to ensure that students entering UAEU have sufficient skills to succeed in the faculty programs, all students entering UAEU must pass three levels of English as a Second Language (ESL) instruction plus an English benchmark exam, currently either TOEFL (Test of English as a Foreign Language) or IELTS (International English Language Testing System), as well as two levels of Arabic, Mathematics and Information Technology. A student can "challenge through" one or more levels either by obtaining sufficiently high grades during placement exams, or, if they achieve an "A" or better in a lower level course, a challenge exam at the end of a semester. A student may complete their English requirement by achieving a sufficiently high standard in a TOEFL or IELTS exam administered by a third party.

Each writing teachers' load is four classes which meet twice a week for 2 hours a day. A writing teacher may see between 60 and 100 students each week. This workload generates a large quantity of marking for each writing teacher, especially since any formal assessment receives double blind marking with a possible third marking if arbitration is required. It might be desirable to have at least one marking of the formal assessments computer-graded. Possibly computer-grading could be useful for other assignments as well.

Commercial packages have recently become available to perform automated grading of writing assignments. Proprietary systems are being used to replace one marker for high-stakes tests such as GMAT® (Graduate Management Admission Test®) and TOEFL® (Test of English as a Foreign Language™)[1] Initial offerings available to an institution such as UGRU tended to be expensive and require a significant amount of "training" making them more suitable for higher volume assessments [2]. Since the start of this project, the writing department has initiated an evaluation of ETS (Educational Testing Service) Criterion, a web-based system that can immediately provide a holistic score and writing analysis feedback to the student [3]. This product does not require specific training and per student pricing is competitive with text book pricing.

This project evaluates the applicability of Classification Association Rules Mining (CARM) algorithms to the problem of computer grading of essays. CARM does not appear to have been applied to the problem in the past, yet it would appear to provide a good fit. The aim of the project is to use CARM to see if we can emulate a "coarse" overall grade for an essay that is acceptably close to what a human rater would assign. If the evaluation proves that the algorithms are applicable, further development could be done towards a more fine-grained grading system that would provide more useful feedback to the student.

The general approach to the problem is to use as much "off-the-shelf" software as possible. Some custom-developed software is required in order to allow teachers to grade essays composed online and to prepare the data for mining. The main effort involves the data collection phase and the pre-processing of the graded writing samples into the appropriate structure required to train and test the mining models.

## 2. EXISTING SCHOLARSHIP

### 2.1 Writing Assessment

In his introduction to the journal, Assessing Writing, Brian Huot outlines the development of scholarship on the assessment of writing. Prior to the 1970's, indirect methods of assessment were most prevalent, and are still widely used today. Huot notes that in 1994, almost 50% of post-secondary institutions still used tests of the mechanics of writing rather than actual writing samples in their first year placement exams. By the end of the 1970's two new styles of assessment prevailed, holistic and primary trait. Holistic grades are arrived at through a general impression of the writing, while primary trait assessment involves examining particular traits that are involved in a specific writing task [4]. A prevalent method of clarifying primary trait grading standards is through use of rubrics. Rubrics define traits that will be graded and qualities exhibited by each grade level. UAE University's University General Requirements Unit (UGRU) uses rubrics in its writing program to assist teachers in grading [5].

Since assessment methods by human raters involve subjective judgements, a key issue is rater reliability or severity [6]. In an institution as large as UGRU, where the same writing test may be taken by over 1,000 students at a time, writing tests are graded double-blind. Where the two marks differ significantly, the grade may be arbitrated by a third marker. More recently, statistical evaluations of rater severity have been applied instead of an arbitrator, leading to situations where the final mark could be higher or lower than the mark assigned by either grader.

Over the past decade a number of systems that automate the grading of essays have been developed. They vary according to their methodology and in the nature of essays that they can grade. This review will outline the current state of the art, discuss the specific issues affecting the grading of Second Language (L2) writing and review applicable data mining technologies.

The development of automated feedback methods for English essays is not new. Computer Aided Instruction (CAI) programs sought to give feedback to students from quite early in the history of computing. Whithaus [7] cites papers from 1962 (Porter) and 1967 (Bloom and Bloom) as examples of early studies into the idea of instant corrective feedback while Williams [8] cites a 1966 paper by E.B. Page, "The imminence of grading essays by computer" as one of the earliest mentions of essay grading by computer. Kukich [9] offered the time line below (Figure 1) for the development of computer writing evaluation (with a more recent focus on research at Educational Testing Service, a private, non-profit organization involved in assessment research, assessment development and test administration [10]).

Currently there are at least ten automated systems for assessment of free text answers available either commercially or as a result of research [11]. A number of different methodologies are represented, the major ones being: the use of proxy characteristics (Project Essay Grade - PEG), semantic analysis (IEA – Intelligent Essay Assessor), natural language processing (Electronic Essay Rater – e-rater), neural networks (Intelligent Essay Marking Systems – IEMS) and text classification (Bayesian Essay Test Scoring System – BETSY) [12] [13]. Systems may be classified by style using a two dimensional taxonomy developed by Page [14]. The first dimension is whether the essay is being graded for writing style or for content. The second dimension is whether the system grades using proxies to simulate raters' criteria or if it grades using some master or "gold standard" as a point of comparison [15]. Valenti et al note that more recent systems tackle both content and style using one of the rating methodologies and classifies these systems as shown in Table 1 below.

Second language writing assessment presents additional difficulties for automated grading. Ferris notes that text analysis of L2 (second language) texts by programs designed for L1 (first language) texts has a high probability of being unable to classify features due to errors by the language learner and argues that analysis programs should be specifically developed to target L2 texts [16]. For very low levels, just understanding what the writer is attempting to communicate can be difficult for a human let alone a computer. At higher levels, a number of features can be found that correlate well with increased proficiency such as longer essays with broader word choice [17].

E-rater has been tested against Test of Written English (TWE) responses for writers in several language groups including Arabic. Good agreement was found between the system and human raters. The features selected by the system to assess the writing were similar to those selected when training with L1 essays. The writing prompt was found to have a significant effect in scoring for Arabic speakers. [18]

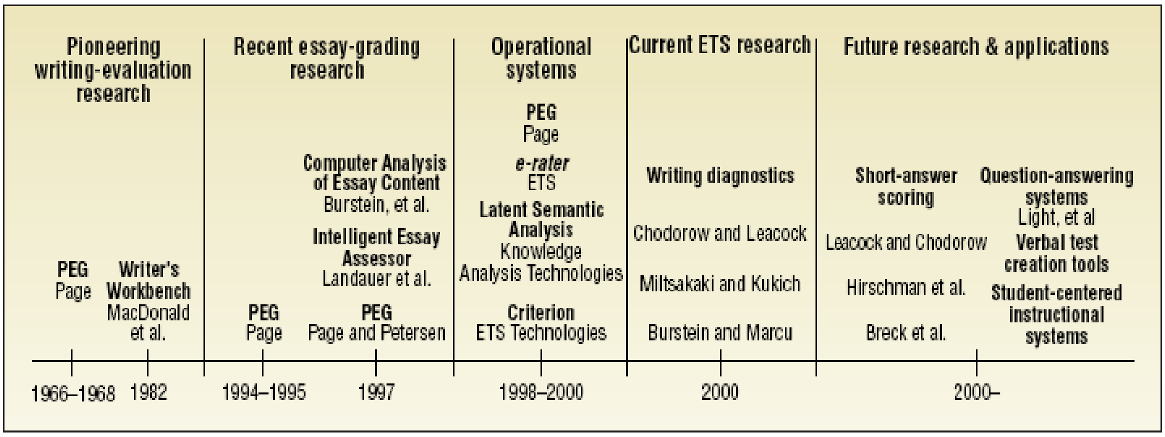

Pioneering writing-evaluation research
Recent essay-grading research
Operational systems
Current ETS research
Future research & applications
PEG Page
Writer's Workbench MacDonald et al.
PEG Page
Computer Analysis of Essay Content Burstein, et al.
Intelligent Essay Assessor Landauer et al.
PEG Page and Petersen
PEG Page
e-rater ETS
Latent Semantic Analysis Knowledge Analysis Technologies
Criterion ETS Technologies
Writing diagnostics
Chodorow and Leacock
Miltsakaki and Kukich
Burstein and Marcu
Short-answer scoring
Leacock and Chodorow
Hirschman et al.
Breck et al.
Question-answering systems Light, et al
Verbal test creation tools
Student-centered instructional systems
1966–1968
1982
1994–1995
1997
1998–2000
2000
2000–

**Fig. 1:** A timeline of research developments in writing evaluation.

**Table 1:** Automated essay grading system's classification.

| | Rating Simulation | Master Analysis |
|---|---|---|
| Content | IEA, BETSY, IEMS, SEAR | ETS I, E-Rater, C-Rater, Automark, PS-ME |
| Style | PEG, BETSY, IEMS, SEAR | E-Rater, Automark, PS-ME |

Meara, Rodgers and Jacobs have applied a neural network to scoring essays in French by non-native speakers. They had good success using sets of ten target words selected from words appearing in 50% of the essays [19]. Lonsdale and Strong-Krause used the cost vector from a slightly modified link grammar parser as a proxy for a human marker. The modifications consisted of adding lexical items common in the students' usage, variant date constructions, allowing for unexpected commas and penalizing some ungrammatical constructions which would be unlikely in L1 writing. They reported that about two thirds of the time the parser agreed with the human raters [20].

### 2.2 Text Data Mining

Text data mining is usually thought to be related to text classification more so than knowledge discovery, which tends to be connoted by "data mining." Computational linguistics is most interested in finding patterns in textual data and is closely related to text data mining [21]. Essay grading is most similar to text classification where a "training set" of classified (graded) text is compared to a "test set" of unclassified or "unlabelled" text in an attempt to classify or "label" the test set.

Numerous algorithms have been proposed for text classification. Naïve Bayes algorithms [22] assume conditional independence of attributes, and classifies documents according to the probability that the words they contain belong to a given class. Support Vector Machines [23] represent a document as a vector and then classify according to the similarity to prototype vectors that are already classified. K-nearest neighbour classification (kNN) finds the k-nearest neighbours among training documents and sums the weights in a category to arrive at a score [24].

Classification Association Rules Mining (CARM) is the application of Association Rule Mining (ARM) technology to the Classification Rule Mining (CRM) problem, in some sense it is therefore a refinement of ARM [25]. ARM seeks to identify hidden patterns, referred to as association Rules (ARs), in data sets comprising only binary valued attributes. CRM seeks to identify and formulate Prediction Rules (PRs) from data sets comprising a variety of attributes. CARM is then the process of finding PRs in binary valued data sets. The discipline of text mining is concerned with the mining of document bases to identify (say) ARs or PRs [26]. Such systems operate by first translating the document base into an appropriate format so that ARM, CRM or CARM techniques can be applied as appropriate.

### 2.3 Conclusion

Researchers have developed a number of commercially viable techniques for automated essay scoring that show a good correlation of results with trained human raters. Each product uses slightly different techniques or blends of techniques to arrive at a score. Scoring essays by second language learners presents a particular challenge to automated scoring due to the likelihood of mechanical errors by the students.

The author has not been able to find evidence that Classification Association Rules Mining has been applied to the problem of automated essay grading. It is suggested here that CARM is ideally suited to the type of analysis required to score an essay.

## 3. EXPERIMENT

This project consists of a number of components, some of which were available "off-the-shelf" and some of which were developed for this project. There are four components required for this project: (i) data gathering, (ii) pre-processing for data mining, (iii) training the data mining model and (iv) validating the trained model against the testing data.

The Level 3 writing program agreed to administer a timed writing information transfer assessment by computer for a portion of their classes. This is a low stakes assessment that does not affect the student's grade much and therefore does not overly penalize students with poor keyboarding skills. Test administration was done using existing CGE (Computer Generated Exam) software widely used in UGRU. The CGE program is locally developed. The essay features of the program had not been used as heavily as the multiple-choice and gap fill question types so the facility to grade the essay responses on line had not yet been created for lecturers use. The author developed a program to allow lecturers to read each response on line and enter a grade and comments for the response into a database. This facilitated easy capture the grading data required for training as well as removing the need for paper copies. Lecturers were used to marking writing on line from the CEPA (Common English Proficiency Assessment) that is administered nation-wide to all high-school students annually. The "portfolio" grade assigned was the target attribute for the CARM algorithm. The portfolio grade is the grade that would contribute to the students writing portfolio mark. It could potentially be in a range from 0 to 3 rather than the detailed grade that could range from 0 to 16.

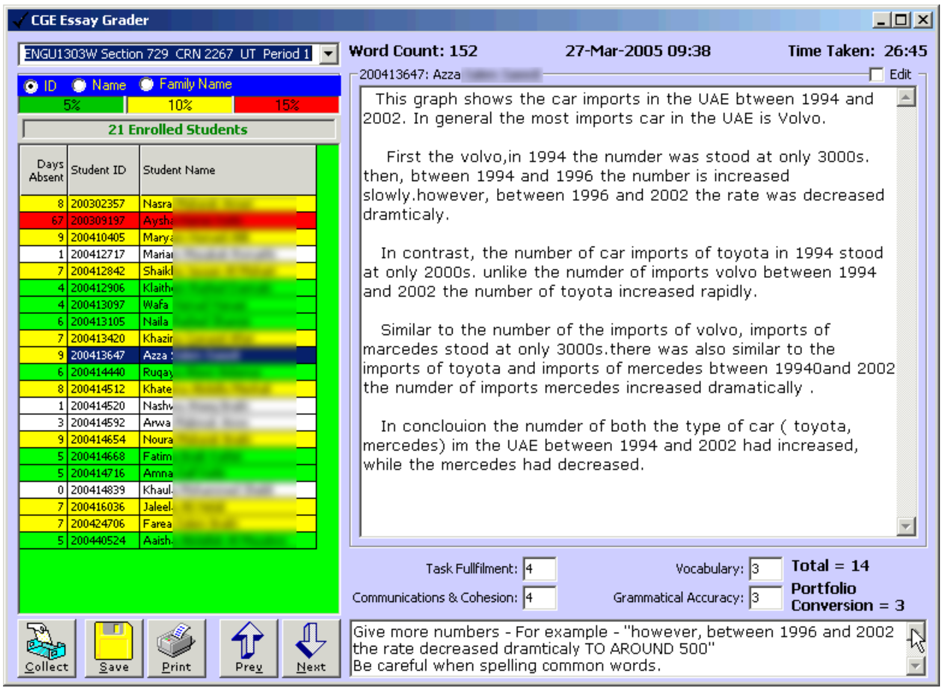
The CGE Essay Grader serves two purposes: (i) to gather the essays and results from the text files and store them in the database and (ii) to allow the instructor to review the essays, grade them and disseminate the results. The data structures required for the essay grader build on those needed to support task preparation. A record is included to record the student's essay and parameters about the exam environment such as student location, time the exam was written and the time the student took to write the exam. A second record is related to each essay record. This record adds the marker's identification to the key allowing an unlimited number of markers to independently grade each essay. The marker may annotate the student's work as well as record detailed grades and any comments for feedback to the students. As far as possible, the program prevents invalid inputs. Figure 2 shows a sample dialogue as used during grading.

The major problem with mining text composed by second language learners is that errors in composition such as lack of white space, misspelled word and inappropriate construction can confuse any automated analysis. A number of steps were taken to solve these issues by introducing appropriate white space and attempting to correct spelling errors prior to parsing and performing linguistic analysis on the text.

The first step was to deal with the problem of white space. The sample essay that has been used to illustrate the user interface, exhibits this issue 4 times ("volvo,in", "slowly.however", "3000s.there", "19940and"). The problem is exacerbated by some valid usages that might normally force white space such as the use of the comma as a thousands separator. The use of "s" at the end of a number to make it plural seems to be common among the group of students, so was left as is. A series of four regular expressions were used to modify the text. The regular expression engine used was VBScript Regular Expressions 5.0.

The second issue to deal with was the problem of spelling errors. Again, the sample essay shows quite a few spelling errors that a human reader could easily handle in context but might prove problematic for a machine reader (e.g. "numder", "btween", "marcedes", "conclouion", "im"). Using an interactive spell-checker would be problematic because of the volume of work. An automated solution is required. The original plan was to use VSSPELL 8.0 from the Component One Studio. Initial investigations appeared somewhat promising as the component would provide an error count and methods for navigating through errors. Further work showed that the control was inappropriate for this application as it was better suited for interactive use in a custom application and the dictionary did not seem comprehensive enough.

The second tool considered was the object model for Microsoft Word 2003. Using Word as a component offered a number of objects and methods that are useful. The ReadabilityStatistics object provides eleven statistics: Words, Characters, Paragraphs, Sentences, Sentences per Paragraph, Words per Sentence, Characters per Word, Passive Sentences, Flesch Reading Ease, and Flesch-Kincaid Grade Level. Readability statistics were added to the mining attribute data table for potential use in the mining algorithm. The SpellingErrors property returns a ProofReadingErrors collection object containing any spelling errors. The SpellingSuggestion object contains a list of possible correct spellings. The count of spelling errors was added to the mining attribute table. Each spelling error was replaced with the first suggestion made by Word's spell checker. In order to aid in evaluation of the spell checking algorithm, a table was maintained with a record for each misspelling with the list of words suggested, the count of words suggested and the replacement chosen. Figure 3 shows a sample essay after spell checking.



CGE Essay Grader

ENGU1303W Section 729 CRN 2267 UT Period 1

● ID ○ Name ○ Family Name

5% 10% 15%

21 Enrolled Students

| Days Absent | Student ID | Student Name |
|---|---|---|
| 8 | 200302357 | Nasra [redacted] |
| 67 | 200309197 | Aysha [redacted] |
| 9 | 200410405 | Marya [redacted] |
| 1 | 200412717 | Mariar [redacted] |
| 7 | 200412842 | Shaikl [redacted] |
| 4 | 200412906 | Klaith [redacted] |
| 4 | 200413097 | Wafa [redacted] |
| 6 | 200413105 | Naila [redacted] |
| 7 | 200413420 | Khazir [redacted] |
| 9 | 200413647 | Azza [redacted] |
| 6 | 200414440 | Ruqay [redacted] |
| 8 | 200414512 | Khate [redacted] |
| 1 | 200414520 | Nashv [redacted] |
| 3 | 200414592 | Arwa [redacted] |
| 9 | 200414654 | Noura [redacted] |
| 5 | 200414668 | Fatim [redacted] |
| 5 | 200414716 | Amna [redacted] |
| 0 | 200414839 | Khaul [redacted] |
| 7 | 200416036 | Jaleel [redacted] |
| 7 | 200424706 | Farea [redacted] |
| 5 | 200440524 | Aaish [redacted] |

Word Count: 152 27-Mar-2005 09:38 Time Taken: 26:45

200413647: Azza [redacted] □ Edit

This graph shows the car imports in the UAE btween 1994 and 2002. In general the most imports car in the UAE is Volvo.

First the volvo,in 1994 the numder was stood at only 3000s. then, btween 1994 and 1996 the number is increased slowly.however, between 1996 and 2002 the rate was decreased dramticaly.

In contrast, the number of car imports of toyota in 1994 stood at only 2000s. unlike the numder of imports volvo between 1994 and 2002 the number of toyota increased rapidly.

Similar to the number of the imports of volvo, imports of marcedes stood at only 3000s.there was also similar to the imports of toyota and imports of mercedes btween 19940and 2002 the numder of imports mercedes increased dramatically .

In conclouion the numder of both the type of car ( toyota, mercedes) im the UAE between 1994 and 2002 had increased, while the mercedes had decreased.

Task Fullfilment: 4
Vocabulary: 3
Total = 14
Communications & Cohesion: 4
Grammatical Accuracy: 3
Portfolio Conversion = 3

Give more numbers - For example - "however, between 1996 and 2002 the rate decreased dramticaly TO AROUND 500"
Be careful when spelling common words.

**Fig. 2.:** CGE essay grader screen

This graph shows the car imports in the UAE between 1994 and 2002. In general the most imports car in the UAE is Volvo.

First the Volvo, in 1994 the number was stood at only 3000s. then, between 1994 and 1996 the number is increased slowly. however, between 1996 and 2002 the rate was decreased dramatically.

In contrast, the number of car imports of Toyota in 1994 stood at only 2000s. unlike the number of imports Volvo between 1994 and 2002 the number of Toyota increased rapidly.

Similar to the number of the imports of Volvo, imports of mar cedes stood at only 3000s. there was also similar to the imports of Toyota and imports of Mercedes between 19940 and 2002 the number of imports Mercedes increased dramatically.

In conclusion the number of both the type of car ( Toyota, Mercedes) imp the UAE between 1994 and 2002 had increased, while the Mercedes had decreased.
Two products were available for parsing and surface linguistic analysis, both by Connexor Oy: Machinese Phrase Tagger and Machinese Syntax. Both products parse the text and perform morphological and syntactic

analysis, but are implemented differently and Syntax provides deeper syntactic analysis. Both products were evaluated using the sample Visual Basic COM applications provided by Connexor for the Windows operating systems. Phrase Tagger operates more as a standard COM component providing an easily accessible set of properties for each word parsed. Syntax returns an XML document or text document for the entire analysis which would need to be re-parsed into a format that could be consumed by the database. Given its better fit for this application and the perceived lack of a requirement for a deep level of syntactic analysis, Phrase Tagger was chosen as the parser and linguistic analyzer. Phrase Tagger was used to feed an intermediate table to hold the parsed text and linguistic information for each essay.

Since Ten Cross Validation (TCV) was being used to conduct the actual mining experiment, each essay needed to be assigned to a scenario. This was done by ordering the essays by grade then assigning each a number from 0 through 9 that could be used as criteria by SQL views to include the essay in either a training or a testing group.

At this stage, it was possible to prepare the data for the mining model. Each unique word was added to an intermediate table together with its number of occurrences and the number of essays it appeared in. The word identifiers were added to the mining attributes for each essay with the exception of eight stop words, mainly those that appeared in the prompt. The stop words were: “and”, “in”, “Mercedes”, “the”, “to”, “Toyota”, “UAE”, and “Volvo”. Initially a much longer stop word list was used, but it seemed to be too draconian, making the resulting text vector very short. In addition to the text vector, other surface attributes were available for data mining. The mining algorithm requires attributes to be discrete. A stored procedure was used to split the attributes in to appropriate ranges.

The Microsoft Association Rules Mining algorithm in SQL Server 2005 has a user-friendly interface that makes it relatively easy to set up and run a mining model. A set of SQL Server views were created for each scenario of training and testing data. For each experiment there is a training case view, a training attribute view, a test case view and a test attribute view. A mining model was created for each experiment consisting of a training case view and a nested attribute table. Setting the predicted column to be the overall “Portfolio Grade” allows the algorithm to act as a Classification Association Rules Mining algorithm restricting rules generated to ones that predict that attribute.

The algorithm provides a number of parameters that can be varied when running the model. The key variables are the minimum and maximum support, the minimum and maximum probability (confidence level) and the minimum and maximum item set size.

## 4. EVALUATION OF DATA MINING

In general, the main issues with data mining were the marginal number of essays and the lack of differentiation between the grade levels for most attributes. When running mining models, the Microsoft Association Rules Algorithm reported “marginal model” frequently. Both IEA [27] and ETS [28] report the requirement for 200 or more essays for training using graded student responses only. The large number of essays with a grade of 2 as compared to 1 or 3 made it more likely that as minimum support was raised, rules for only grade 2 would be generated. Unless otherwise specified, the tables shown in the following sections are for mining models with a confidence parameter of 50% and a minimum support level of 6 items.

When only the text vector was used for data mining, no test cases received a grade of one for any scenario. Sensitivity tests for confidence level were the most stable of the three sets of attributes used. The high confidence rules tended to correctly classify grades 2 and 3. Lower confidence rules classified grade 1 essays as 2 or 3, while the higher confidence rules classified all grade 1 essays as 2, which is preferable to classifying them as grade 3. When the support was varied, having a support of 4 produced rules that tended to classify just about all essays as grade 3. As support increased, rules rapidly converged to assign only a grade of 2. By the time support reaches 10, the vocabulary available for mining has been reduced to 90 words which would greatly shorten the text available for mining. (Table 2).

Using a confidence level of 90% and minimum support of 6, results were quite reasonable in the TCV tests. All scenarios had a better than 50% accuracy rate with a maximum of 75%. Using a confidence level of 50% produced too many grade 3 rules leading to misclassifications in most scenarios. (Table 3).

Unlike the text only scenario, use of surface linguistic attributes only failed to correctly classify the grade 3 essays although it did classify a few essays as grade 3. We see the same pattern in the sensitivity analysis as support increases, accuracy increases, largely due to the greater likelihood of rules assigning grade 2.

However in this case even with a support level of 16 some grade 1 essays were being classified correctly. This was due to the word count attribute. As confidence increased, so did classification accuracy, although the under 60% accuracy is marginal. (Table 4)

Mining accuracy is quite good at a 90% confidence level. Two scenarios hit 80% classifying 3 of 4 grade 1 essays and 9 of 10 grade 2 essays correctly. At a 90% classification level, results look better than for text only. The 50% confidence level produced poor results, in no case exceeding 50% accuracy. (Table 5)

Combining both text and surface linguistic attributes produced the only scenario where at least some of all three grade levels were correctly classified. Similar patterns emerge in the sensitivity analysis with improving accuracy with increasing confidence and support. The scenarios with poor results tended to over-classify the grade 2 essays as grade 3 or grade 1. (Table 6)

This set of attributes had the most variability in classification accuracy of the three used. The best case was 75% accuracy while the worst case was 33%. When used with either set of attributes uncombined, the worst case at 90% confidence was better than 50% accuracy. At a 50% confidence level a number of scenarios grossly over-classified the essays, assigning a grade of 3 to almost all essays. (Table 7)

**Table 2:** Sensitivity analysis (text only).

| Confidence | Correct (of 16) | Percent correct |
|---|---|---|
| 100% | 12 | 75.00% |
| 90% | 12 | 75.00% |
| 80% | 11 | 68.75% |
| 70% | 9 | 56.25% |
| 60% | 9 | 56.25% |
| 50% | 9 | 56.25% |
| 40% | 9 | 56.25% |
| 30% | 9 | 56.25% |
| 20% | 9 | 56.25% |
| 10% | 9 | 56.25% |

| Minimum Support | Correct (of 16) | Percent Correct |
|---|---|---|
| 2 | 7 | 43.75% |
| 4 | 3 | 18.75% |
| 6 | 9 | 56.25% |
| 8 | 6 | 37.50% |
| 10 | 10 | 62.50% |
| 12 | 10 | 62.50% |
| 14 | 10 | 62.50% |
| 16 | 10 | 62.50% |

**Table 3:** Mining accuracy (text only).

| Scenario | Confidence = 50% | | Confidence = 90% | |
|---|---|---|---|---|
| | Correct | Percent Correct | Correct | Percent Correct |
| 0 | 9 | 56.25% | 12 | 75.00% |
| 1 | 6 | 37.50% | 9 | 56.25% |
| 2 | 5 | 31.25% | 10 | 62.50% |
| 3 | 3 | 20.00% | 11 | 73.33% |
| 4 | 8 | 53.33% | 11 | 73.33% |
| 5 | 3 | 20.00% | 9 | 60.00% |
| 6 | 5 | 33.33% | 8 | 53.33% |
| 7 | 6 | 40.00% | 8 | 53.33% |
| 8 | 2 | 13.33% | 10 | 66.67% |
| 9 | 3 | 20.00% | 9 | 60.00% |

**Table 4:** Sensitivity analysis (surface linguistic attributes).

| Confidence | Correct (of 16) | Percent correct |
|---|---|---|
| 100% | 9 | 56.25% |
| 90% | 9 | 56.25% |
| 80% | 7 | 43.75% |
| 70% | 5 | 31.25% |
| 60% | 5 | 31.25% |
| 50% | 6 | 37.50% |
| 40% | 6 | 37.50% |
| 30% | 6 | 37.50% |
| 20% | 6 | 37.50% |
| 10% | 6 | 37.50% |

| Minimum Support | Correct (of 16) | Percent Correct |
|---|---|---|
| 2 | 2 | 12.50% |
| 4 | 4 | 25.00% |
| 6 | 6 | 37.50% |
| 8 | 6 | 37.50% |
| 10 | 10 | 62.50% |
| 12 | 11 | 68.75% |
| 14 | 12 | 75.00% |
| 16 | 12 | 75.00% |

**Table 5:** Mining accuracy (surface linguistic attributes).

| | Confidence = 50% | | Confidence = 90% | |
|---|---|---|---|---|
| Scenario | Correct | Percent Correct | Correct | Percent Correct |
| 0 | 6 | 37.50% | 9 | 56.25% |
| 1 | 4 | 25.00% | 11 | 68.75% |
| 2 | 6 | 37.50% | 11 | 68.75% |
| 3 | 3 | 20.00% | 9 | 60.00% |
| 4 | 7 | 46.67% | 11 | 73.33% |
| 5 | 7 | 46.67% | 10 | 66.67% |
| 6 | 6 | 40.00% | 8 | 53.33% |
| 7 | 6 | 40.00% | 9 | 60.00% |
| 8 | 6 | 40.00% | 12 | 80.00% |
| 9 | 5 | 33.33% | 12 | 80.00% |

**Table 6:** Sensitivity analysis (combined).

| Confidence | Correct (of 16) | Percent correct |
|---|---|---|
| 100% | 10 | 62.50% |
| 90% | 10 | 62.50% |
| 80% | 8 | 50.00% |
| 70% | 5 | 31.25% |
| 60% | 4 | 25.00% |
| 50% | 4 | 25.00% |
| 40% | 4 | 25.00% |
| 30% | 4 | 25.00% |
| 20% | 4 | 25.00% |
| 10% | 4 | 25.00% |

| Minimum Support | Correct (of 16) | Percent Correct |
|---|---|---|
| 2 | 8 | 50.00% |
| 4 | 4 | 25.00% |
| 6 | 4 | 25.00% |
| 8 | 7 | 43.75% |
| 10 | 9 | 56.25% |
| 12 | 10 | 62.50% |
| 14 | 12 | 75.00% |
| 16 | 10 | 62.50% |

**Table 7:** Mining accuracy (combined).

| | Confidence = 50% | | Confidence = 90% | |
|---|---|---|---|---|
| Scenario | Correct | Percent Correct | Correct | Percent Correct |
| 0 | 4 | 25.00% | 10 | 62.50% |
| 1 | 6 | 37.50% | 12 | 75.00% |
| 2 | 3 | 18.75% | 8 | 50.00% |
| 3 | 3 | 20.00% | 10 | 66.67% |
| 4 | 8 | 53.33% | 9 | 60.00% |
| 5 | 8 | 53.33% | 11 | 73.33% |
| 6 | 2 | 13.33% | 5 | 33.33% |
| 7 | 6 | 40.00% | 8 | 53.33% |
| 8 | 3 | 20.00% | 11 | 73.33% |
| 9 | 4 | 26.67% | 8 | 53.33% |

## CONCLUSION

The biggest issue for this project was the lack of sample essays. We were surprised that a number of lecturers who are proponents of writing on computer did not participate in the data collection phase. The decision to gather essays using the information transfer timed writing was taken after the semester had started. It is likely that these lecturers were handling the task earlier in their course pacing and therefore were not able to accommodate this experiment. Any continuation of this experiment will require a decision on the target prompt prior to the beginning of the semester and a mandate from the program coordinator for lecturers to participate.

Given the type of writing elicited from the students on the information transfer task, it may be that it is not ideally suited for the experiment given that writing was so repetitive.

A larger volume of samples would likely improve the model, especially if the numbers of essays in each grade band can be levelled out to allow better support for rules defining the grades at the outer ends of the spectrum. Choosing a different writing type may also improve results. The information transfer task elicited a restricted vocabulary while an opinion piece might provide a broader vocabulary for text mining.

In the past year, keyboarding skills have received a greater emphasis in the first-year curriculum. A keyboarding test has been added to the Information Technology course's evaluation criteria. Improved student keyboarding skills will make collection of writing samples using computer more viable and valid.

A key issue in training the mining model is rater consistency and severity. This investigation used a single rater to assign a grade to each essay. In reviewing the essays, the author noted a number of instances where there were likely inconsistencies in grading and departures from the scoring rubric. For several terms, UGRU's assessment coordinator has been applying a Rasch model to mitigate against rater inconsistency and rater severity variations [29] in double-blind marking of higher stakes writing assignments. The Rasch model depends on an appropriate network of graders having marked the same items. The model then fits the statistics to arrive at independent estimates of student and rater performance. It can then identify occurrences such as a consistent lenient rater and a consistent severe rater may have differed on the mark assigned by an amount that would normally have sent the paper to arbitration but really doesn't require it versus the case where two grades were the same and one was by a severe rater and the other by a lenient rater and therefore perhaps that case should be arbitrated. It is likely that smoothing out grading inconsistencies would lead to better mining model training results.

Additional investigation could be made into two different aspects of the attributes used for data mining. This investigation employed the actual words used by the student (as modified for spelling errors). Connexor Machinese Phrase Tagger also makes available the base form of the word. Using the base form may provide different results from text mining. The other aspect that could be investigated is deriving further attributes from the surface linguistic characteristics. Some potential attributes are nouns per sentence, verbs per sentence or number of sentences without a subject, object and verb.

The results of this investigation were marginal due to small number of samples, the repetitive nature of the writing and rater inconsistencies. Evidence from the best case mining models indicates that future work in applying CARM to essay grading is worth pursuing. CARM is closely related to the Bayesian approaches taken by others. Use of both all three approaches (text only, surface linguistic attributes and the two combined) showed promising results. It is likely that with a better distribution of training data, useful results would emerge. The tools provided by Microsoft's forthcoming data mining suite make CARM very accessible. While Microsoft does not label their package as a CARM tool, using their ARM algorithm to predict a single attribute effectively makes it one. It will be interesting to see how a larger sample works with these tools.

## ACKNOWLEDGEMENT

This work was originally prepared in fulfillment of the dissertation requirement of the MSc in Information Technology, University of Liverpool. Special thanks to Dr. Frans Coenen, my dissertation advisor, and Dr. Abdullah Al Khanbashi for sponsoring the project. Thanks to the UGRU writing staff coordinated by Ms. Jodi Lefort for their assistance in gathering the data and testing the online grading components. The granting of a thesis license by Connexor Oy for their linguistic software greatly broadened the scope of this investigation.

## REFERENCES

[1] ETS, "ETS Launches New Certified Education Partner Program To Deliver Next Generation TOEFL Test" [Internet]. http://www.ets.org/news/04110201.html. Princeton, N.J. (Updated: Nov. 2, 2004) 2004.

[2] Palmer, J., Williams, R. and Dreher, H. "Automated Essay Grading System Applied to a First Year University Subject – How Can We do it Better?" Informing Science, June 2002 [Online]. Available from: http://proceedings.informingscience.org/IS2002Proceedings/papers/Palme026Autom.pdf (Accessed: 30 December 2004).

[3] ETS, "ETS Criterion: Diagnostic grading for testing English language learning from ETS: Description." [Internet]. Available from: http://www.ets.org/criterion/ell/description.html (Updated: 26 April 2005)

[4] Huot, B. "An Introduction to *Assessing Writing*." Assessing Writing, 1 (1), 1-9, 1994.

[5] Coakley, J. and LeFort, J. Level 3 Information Transfer Scoring Criteria, UAE University internal document. 2005.

[6] Gamaroff, R. "Rater reliability in language assessment: the bug of all bears", System, Volume 28, Issue 1, March 2000, pp. 31-53 [Online]. DOI:10.1016/S0346-251X(99)00059-7 (Posted online: 24 January 2000)

[7] Whithaus, C. '"he Development of Early Computer-Assisted Writing Instruction (1960-1978): The Double Logic of Media and Tools", Computers and the Humanities, Volume 38, Issue 2, May 2004. pp. 149 – 162 [Online]. DOI: 10.1023/B:CHUM.0000031171.79841.02 (Accessed: 8 July 2005)

[8] Williams, R. "Automated essay grading: An evaluation of four conceptual models", A. Herrmann and M. M. Kulski (Eds), Expanding Horizons in Teaching and Learning. Proceedings of the 10th Annual Teaching Learning Forum, 7-9 February 2001. Perth: Curtin University of Technology. [Online]. Available from: http://lsn.curtin.edu.au/tlf/tlf2001/williams.html (Accessed: 26 December 2004)

[9] Kukich, K. "Beyond automated essay scoring" IEEE Intelligent Systems 15.5, pp. 22-27, 2000. [Online]. DOI: 10.1109/5254.889104 (Posted online: 6 August 2002)

[10] ETS. "What We Do" [Internet]. http://www.ets.org/aboutets/wedo.html (Updated: July 26, 2004)

[11] Valenti, S., Neri, F. and Cucchiarelli. "An Overview of Current Research on Automated Essay Grading", Journal of Information Technology Education, Volume 2, 2003. [Online]. Available from: http://jite.org/documents/Vol2/v2p319-330-30.pdf (Accessed: 16 December 2004)

[12] Callear, D., Jerrams-Smith, J. and Victor, S. "Bridging gaps in computerised assessment of texts", Proceedings. IEEE International Conference on Advanced Learning Technologies, 2001, pp.139-140 [Online]. DOI: 10.1109/ICALT.2001.943881 (Posted online: 7 August 2002)

[13] Potter, A. 'Invoking the cyber-muse: automatic essay assessment in the online learning environment', P. Baumgartner, P. A. Cairns, M. Kolhase & E. Melis (Eds.), Eighteenth International Joint Conference on Artificial Intelligence Workshop Program: Knowledge Representation and Automated Reasoning for E-Learning Systems, Acapulco, Mexico, 2003, pp.

59-62 [Online]. Available from: http://home.hiwaay.net/~anpotter/PotterA04.pdf (Accessed: 23 December 2004)

[14] Page, E. B. The Imminence of Grading Essays by Computer. Phi Delta Kappan, 48:238-243. 1966.

[15] Williams, R. "Automated essay grading: An evaluation of four conceptual models", A. Herrmann and M. M. Kulski (Eds), Expanding Horizons in Teaching and Learning. Proceedings of the 10th Annual Teaching Learning Forum, 7-9 February 2001. Perth: Curtin University of Technology. [Online]. Available from: http://lsn.curtin.edu.au/tlf/tlf2001/williams.html (Accessed: 26 December 2004)

[16] Ferris, D. "The design of an automatic analysis program for L2 text research: Necessity and feasibility" Journal of Second Language Writing, Volume 2, Issue 2, May 1993, pp. 119-129 [Online]. DOI: 10.1016/1060-3743(93)90013-S (Posted online: 20 June 2002)

[17] Grant, L. and Ginther, A. "Using Computer-Tagged Linguistic Features to Describe L2 Writing Differences", Journal of Second Language Writing, Volume 9, Issue 2, May 2000, pp. 123-145 [Online]. DOI:10.1016/S1060-3743(00)00019-9 (Posted online: 6 November 2000)

[18] Burstein, J., & Chodorow, M. "Automated essay scoring for nonnative English speakers", Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, College Park, MD. June, 1999. [Online]. Available from: http://acl.ldc.upenn.edu/W/W99/W99-0411.pdf (Accessed: 8 July 2005)

[19] Meara, P.M., Rodgers, C., and Jacobs, G. "Vocabulary and neural networks in the computational assessment of texts written by second language learners", System, Volume 28, Number 3, pp. 345-354, 2000 [Online]. DOI: 10.1016/S0346-251X(00)00016-6 (Accessed: 23 December 2004)

[20] Lonsdale, D. & Strong-Krause, D. "Automated Rating of ESL Essays", Proceedings of the HLT-NAACL 2003 Workshop: Building Educational Applications Using Natural Language Processing, 2003 [Online]. Available from: http://acl.ldc.upenn.edu/W/W03/W03-0209.pdf (Accessed on: 9 December 2004)

[21] Hearst, M. "Untangling text data mining", Proceedings of the 37th Annual Meeting of the ACL, pp. 3–10, College Park, Maryland, 1999 [Online]. Available from: http://acl.ldc.upenn.edu//P/P99/P99-1001.pdf (Accessed: 14 January 2005)

[22] Lewis, D. and Gale, W. "A Sequential Algorithm for Training Text Classifiers", W. Bruce Croft and C.J. van Rijsbergen,eds., SIGIR94: Proceedings of Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Springer-Verlag, London, pp.3-12, 1994 [Online]. Available from:
http://santana.uni-muenster.de/Library/Virtual/CorpusLinguistics/lewis.ps (Accessed: 8 July 2005)

[23] Oachims, T. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", European Conference on Machine Learning (ECML), 1998 pp. 137-142 [Online]. Available from: http://www.joachims.org/publications/joachims_98a.ps.gz (Accessed: 8 July 2005)

[24] Yang, Y. and Lui, Y. (1999) "A re-examination of text categorization methods", Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), pp. 42-49 [Online]. Available from: http://www.hpl.hp.com/personal/Carl_Staelin/cs236601/yang1999.ps.gz (Accessed: 8 July 2005)

[25] Liu, B., Hsu, W. and Ma, Y. "Integrating Classification and Association Rule Mining", Proceedings KDD-98, New York, 27-31 August. AAAI. pp. 80-86, 1998 [Online]. Available from: http://citeseer.ist.psu.edu/rd/72594368%2C16476%2C1%2C0.25%2CDownload/http://citeseer.ist.psu.edu/cache/papers/cs/2196/http:zSzzSzwww.comp.nus.edu.sgzSz%7EliubzSzpublicationszSzkdd98_1.ps.gz/liu98integrating.ps.gz (Accessed: 8 July 2005)

[26] Barbara, D., Domeniconi, C. and Kang, N. "Mining relevant text from unlabelled documents", ICDM 2003. Third IEEE International Conference on Data Mining, 2003 [Online]. Available from: http://ieeexplore.ieee.org/iel5/8854/27998/01250959.pdf?isnumber=&arnumber=1250959 (Updated: 19 December 2003)

[27] Landauer, T., Laham, D. and Foltz, P. (2000) 'The Intelligent Essay Assessor', IEEE Intelligent Systems 15, pp. 27-31 [Online]. DOI: 10.1109/5254.889104 (Posted online: 6 August 2002).

[28] Burstein, J., Chodorow, M., & Leacock, C. (2004) 'Automated essay evaluation: The Criterion online writing service', AI Magazine, 25(3), pp. 27-36. [Online]. Available from: http://www.findarticles.com/p/articles/mi_m2483/is_3_25/ai_n6258424 (Accessed: 8 July 2005)

[29] Lunz, M.E., Wright, B.D., Linacre, J.M. "Measuring the Impact of Judge Severity on Examination Scores", Applied Measurement in Education, 3(4), pp. 331-345, 1990 [Online]. Available from: http://www.rasch.org/memo47.htm (Accessed: 8 July 2005).