

Office of Research and Development

**The Theory, Design, Development and Evaluation of the MarkIT
Automated Essay Grading System**

Robert Francis Williams

This thesis is presented for the Degree of

Doctor of Philosophy

of

Curtin University

January 2011

Declaration

To the best of my knowledge and belief this exegesis contains no material previously published by any other person except where due acknowledgment has been made.

This exegesis contains no material, which has been accepted for the award of any other degree or diploma in any university.

Signature: R Williams

Date: 21/11/2011

Table of Contents

The Theory, Design, Development and Evaluation of the MarkIT Automated Essay Grading System	1
Table of Contents	3
Abstract	9
List of Research Publications.....	11
Statement of Contribution by Authors and Others	12
Acknowledgements	15
Chapter 1 Motivation for the Study	16
1.1 Introduction	16
1.2 Preliminary Knowledge about Automated Essay Grading.....	16
1.2.1 An essay.....	16
1.2.2 Essay test.....	17
1.2.3 Essay grading.....	17
1.2.4 Essay grading challenges	17
1.2.5 Computer aided grading system	18
1.2.6 Automated essay grading (AEG) systems	18
1.3 Problems Associated with Automated Essay Grading Systems	19
1.3.1 Summative vs. formative assessment	19
1.3.2 Non-interactive feedback.....	19
1.3.3 Limited computation power	19
1.3.4 Fooling AEG systems	20
1.4 Motivation for the Study.....	20
1.4.1 The technology challenges	20
1.4.2 The cost to governments.....	21
1.4.3 The effort made by teachers	21
1.4.4 The value of feedback	21
1.4.5 Technology adoption in the education sector.....	21

1.5	<i>The Objectives of the Exegesis</i>	23
1.6	<i>Significance of the Research</i>	23
1.7	<i>Overview of the Exegesis</i>	24
1.8	<i>Summary</i>	25
Chapter 2	Literature Review	26
2.1	<i>Introduction</i>	26
2.2	<i>Preliminary Concepts</i>	26
2.2.1	The “Style” in the essay.....	26
2.2.2	The “Content” in the essay.....	26
2.2.3	The “Systems”	26
2.2.4	The scientific approaches that underpin the “Systems”	26
2.2.5	Rubric	27
2.3	<i>Overview of Scientific Approaches from the Literature</i>	27
2.4	<i>Statistical Based Approaches</i>	28
2.4.1	Project Essay Grade (PEG)	28
2.4.2	Paperless School Free Text Marking Engine.....	29
2.5	<i>Latent Semantic Analysis Based Approaches</i>	29
2.5.1	Latent Semantic Analysis (LSA).....	30
2.5.2	Intelligent Essay Assessor (IEA)	30
2.6	<i>Natural Language Processing Based Approaches</i>	30
2.6.2	Educational Testing Service 1 (ETS 1)	31
2.6.3	Electronic Essay Rater (E-Rater)	31
2.6.4	Conceptual Rater (C-Rater)	32
2.6.5	AutoMark	32
2.6.6	Schema Extract Analyse and Report (SEAR)	33
2.6.7	Intellimetric	34
2.7	<i>Bayesian Based Approaches</i>	35
2.7.1	Bayesian Essay Test Scoring System (BETSY).....	35
2.7.2	Text Categorisation Technique (TCT)	35

2.8	<i>Neural Networks Based Approaches</i>	36
2.8.1	Neural networks	36
2.8.2	Intelligent Essay Marking System (IEMS)	36
2.9	<i>Semantic Based Approaches</i>	37
2.9.1	Semantic network	37
2.9.2	SAGrader	38
2.10	<i>Phrase Processing and Term Vector Technologies</i>	39
2.10.1	Phrase processing.....	39
2.10.2	Salton's phrase processing theory	39
2.10.3	NLP phrase processing	39
2.10.4	Term vector	40
2.10.5	Term vector theory.....	41
2.11	<i>Comparison of Scientific Methods Used</i>	42
2.12	<i>A Critical Review of the Existing Approaches to Building AEG Systems.</i>	43
2.13	<i>Summary</i>	44
Chapter 3	Problem Definition	46
3.1	<i>Introduction</i>	46
3.2	<i>Concepts and Definitions</i>	46
3.2.1	The features of an essay.....	46
3.2.2	National literacy and numeracy testing	46
3.3	<i>Literacy and Numeracy Assessment</i>	46
3.3.1	WALNA marking criteria	47
3.3.2	NAPLAN marking criteria	47
3.3.3	NAEP marking criteria	47
3.3.4	The common features of marking criteria (Rubrics)	52
3.4	<i>Challenges for Automated Assessment</i>	53
3.5	<i>Challenges for Automated Literacy Assessment</i>	54
3.5.1	Stylistic aspects	54
3.5.2	Content aspects.....	55

3.5.3 Structure and organization.....	55
3.5.4 Mechanistic aspects	55
3.6 <i>Research Questions</i>	55
3.7 <i>Research Issues</i>	55
3.7.1 Practical limitations of Context Free Phrase Structure Parsers.....	56
3.7.2 Efficient parsing of sentences	56
3.7.3 Structures to hold semantic details.....	57
3.8.1 The methodology used in this research project	58
3.8.2 Nunamaker's, Chen's, and Purdin's methodology	59
3.8.3 Galliers and Land's traditional empirical approaches to IS research	61
3.9 <i>Choice of Appropriate Methodology</i>	62
3.10 <i>Summary</i>	62
Chapter 4 Theoretical Framework of the MarkIT System	64
4.1 <i>Introduction</i>	64
4.2 <i>The Proposed Solutions</i>	64
4.3 <i>An Overview of the Theoretical Framework for MarkIT</i>	66
4.4 <i>Theory of Document Semantic Representation</i>	66
4.4.1 Phrase and clause representations	67
4.4.2 Normalised Word Vector.....	69
4.5 <i>The Theoretical Accomplishment of MarkIT</i>	72
4.6 <i>Summary</i>	72
Chapter 5 The Design and Implementation of MarkIT.....	73
5.1 <i>Introduction</i>	73
5.2 <i>Architectural View of MarkIT</i>	73
5.3 <i>Implementation of the Theoretical Foundation for MarkIT</i>	74
5.4 <i>Essay Feedback Design for MarkIT</i>	79
5.5 <i>Visual Interactive Feedback with MarkIT</i>	80
5.6 <i>Summary</i>	84
Chapter 6 Trial, Testing and Evaluation.....	85

6.1	<i>Introduction</i>	85
6.2	<i>Preparations for Automated Grading with MarkIT</i>	85
6.3	<i>A Trial of the Intelligent Essay Assessor</i>	86
6.4	<i>Evaluation of the Performance of MarkIT</i>	86
6.5	<i>Testing Outcomes</i>	90
6.6	<i>Challenges Encountered in Testing</i>	90
6.7	<i>Summary</i>	91
Chapter 7 Publication of Scientific Papers		92
7.1	<i>Introduction</i>	92
7.1.1	Automated Essay Grading: An Evaluation of Four Conceptual Models	92
7.1.2	Automatically Grading Essays with MarkIT	92
7.1.3	Telecommunications Use in Education to Provide Interactive Visual Feedback on Automatically Graded Essays	92
7.1.4	Formative Assessment Visual Feedback in Computer Graded Essays	93
7.1.5	MarkIT in Action - A Report on the Automatic Grading of DET Year 10 Essays	93
7.1.6	The Power of Normalised Word Vectors for Automatically Grading Essays	93
7.1.7	A Computational Effective Document Semantic Representation	94
Chapter 8 Recapitulation and Future Work		95
8.1	<i>Introduction</i>	95
8.2	<i>Overview of the Research</i>	95
8.3	<i>Contribution of the Research</i>	95
8.3.1	Scientific contribution	95
8.3.2	Contribution to the education sector and community	95
8.3.3	Achieving correlation of 0.81 with human markers	96
8.3.4	Integrated summative and formative assessment	96
8.3.5	Feedback to both the teachers and students	96
8.3.6	Dynamic and interactive feedback	96
8.3.7	Contribution to smart information use as designated by ARC	97
8.3.8	A guide book on technology adoption for the education sector	97

8.3.9 Enhanced teaching and learning experience for both teachers and students.....	97
8.3.10 Economic benefit to the Government and education sector.....	97
8.3.11 Just in time quality feedback.....	98
8.4 <i>Limitations of the Research</i>	98
8.5 <i>Future Directions</i>	99
References.....	100
Glossary of Terms	105
Appendix A Statements of Co-Authorship	108
Appendix B Copyright Permissions	110
Appendix C Publications and Bibliography.....	126

Abstract

The research presented in this exegesis relates to the design, development and testing of a new Automated Essay Grading (AEG) system. AEG systems make use of Information Technology (IT) to grade essays. The major objective for AEG system developers is to build systems that grade as well as, or exceed the accuracy of, human graders.

This research discusses the main theories that currently underpin existing systems. It then discusses a new theoretical concept, the Normalised Word Vector (NWV), which has been developed and tested during this research. This exegesis also synthesises into a cohesive discourse seven of the author's papers on the NWV and related issues published during the period 2002 to 2007. The papers can be grouped into three themes as follows:

- the theory of NWV and related matters,
- the development of the system, and
- the testing of the system.

The exegesis is structured around these themes. A prototype program, named MarkIT, which has been developed and tested using this theory, is also discussed.

Thirteen existing AEG systems have been identified in this research. Each system has its own set of unique features; some focus on grading for essay writing style, others for essay content, and others attempt to consider both aspects in assigning a score to an essay. The type and amount of feedback on an essay also varies amongst the systems; some provide feedback on essay mechanics and others provide feedback on missing content. The MarkIT system described in this exegesis primarily grades for essay content, with a secondary focus on style. It has the unique feature, which distinguishes it from the other systems, of providing interactive visual feedback on essay content. This enables the teacher and student to discuss how the essay can be improved to obtain a higher grade.

In brief, the theory of the NWV is as follows. The words in an essay are 'normalised' to their root concepts in a thesaurus. The number of times these concepts occur in the essay (the counts) are then used to build the coordinates of the vector in the vector space induced by all the concepts in the thesaurus. This adaptation of the theory used for many years in the document retrieval industry enables very fast comparison of essay content, and enables MarkIT to grade in real time.

In essence the system works by mathematically modelling, using multiple linear regression, the grading criteria used by human graders for a given essay. These criteria are extracted from a set of training essays, and include items such as the number of words, the number of nouns, the number of verbs, the number of adjectives, and the number of adverbs. The model is then used to grade the essays not previously graded by humans. It does this by measuring the predictor factors in the ungraded essays, and then applying the multiple regression equation. The cosine of the angle between the NWV for a student essay and the NWV for a model answer is often one of the significant predictor variables.

The system has been tested with 390 Year 10 high school essays, of about 400 words in length, on the topic of 'The School Leaving Age'. The correlation of grades amongst the human graders was 0.81, and the system scores matched this correlation of the human graders.

List of Research Publications

Paper	Chapter discussed
Williams, R (2001), Automated Essay Grading: An Evaluation of Four Conceptual Models, in Kulski, M & Herrmann, A (eds.), <i>New Horizons in University Teaching and Learning: Responding to Change</i> , Curtin University, Perth, Australia.	2
Williams, R & Dreher, H (2004), Automatically Grading Essays with MarkIT, <i>Journal of Issues in Informing Science and Information Technology</i> , vol. 1, pp. 693-700.	5
Williams, R & Dreher, H (2005), Telecommunications Use in Education to Provide Interactive Visual Feedback on Automatically Graded Essays, <i>Proceedings of International Telecommunications Society Africa-Asia-Australasia Regional Conference</i> , Perth, Australia.	4
Williams, R & Dreher, H (2005), Formative Assessment Visual Feedback in Computer Graded Essays, <i>Journal of Issues in Informing Science and Information Technology</i> , vol. 2, pp. 23-32.	5
Williams, R (2005), MarkIT in Action - A Report on the Automatic Grading of DET Year 10 Essays, Curtin University, Perth, Australia.	6
Williams, R (2006), The Power of Normalised Word Vectors for Automatically Grading Essays, <i>Journal of Issues in Informing Science and Information Technology</i> , vol. 3, pp. 721-729.	4
Williams, R (2007), A Computational Effective Document Semantic Representation, <i>Proceedings of IEEE-Digital Ecosystems and Technologies 2007 Conference</i> , Cairns, Australia, pp. 410-415.	4

Statement of Contribution by Authors and Others

I have published 7 papers related to this exegesis. Three of these are journal papers that were initially presented at conferences. Three others are conference papers. The final paper is a report produced for the Western Australian Department of Education and Training (WADET). Papers 2-4 are co-authored papers, and the others are single author papers written by myself. All papers, except paper 5, are qualified and listed in the Curtin Institutional Repository known as the Script Database for RPI (Research Performance Index) collection and can be searched in the open space known as Curtin E-Space (<http://espace.library.curtin.edu.au>). All papers, except paper 5, are refereed and these can be downloaded from the Curtin Script Database. These papers are included in Appendix C of this exegesis. Detailed breakdowns of co-authorship contributions are provided in appendix A.

Paper 1.

Williams, R (2001), Automated Essay Grading: An Evaluation of Four Conceptual Models, in Kulski, M & Herrmann, A (eds), *New Horizons in University Teaching and Learning: Responding to Change*, Curtin University, Perth, Australia.

I was the sole author of this refereed paper. This paper discussed four existing AEG systems, and described their features and performance. This paper emanated from an interest in AEG systems that led the author to survey existing AEG systems in order to understand how they worked, and to ascertain their strengths and weaknesses. The paper also discussed a trial of the Intelligent Essay Assessor conducted at Curtin University. This paper was adjudged a best paper at the 2001 Teaching and Learning Forum (TLF 2001).

Paper 2.

Williams, R & Dreher, H (2004), Automatically Grading Essays with MarkIT, *Journal of Issues in Informing Science and Information Technology*, vol. 1, pp. 693-700.

I was a co-author of this refereed paper. I wrote the section which outlines the features of MarkIT. I also co-authored the section on the performance of MarkIT. The paper discussed an early version of the MarkIT AEG system, and some results obtained with testing it with a small number of university law essays.

Paper 3.

Williams, R & Dreher, H (2005), Telecommunications Use in Education to Provide Interactive Visual Feedback on Automatically Graded Essays, *Proceedings of International Telecommunications Society Africa-Asia-Australasia Regional Conference*, Perth, Australia.

I was a co-author of this refereed paper. I wrote the sections on the theory of semantic representations and the section on the comparison of MarkIT and the human scores. This paper discussed the MarkIT AEG system with particular emphasis on the interactive visual feedback mechanisms. These mechanisms were made possible by the unique way the system represents essay concepts.

Paper 4.

Williams, R & Dreher, H (2005), Formative Assessment Visual Feedback in Computer Graded Essays, *Journal of Issues in Informing Science and Information Technology*, vol. 2, pp. 23-32.

I was a co-author of this refereed paper. I researched and wrote the section of the paper which discusses the production AEG systems. I also wrote the section detailing the features of MarkIT. The paper discussed the MarkIT system feedback mechanisms in the light of formative assessment. As the research proceeded with the new system, new aspects of the system were explored and documented. The system proved to be useful in providing formative feedback to teachers and students.

Paper 5.

Williams, R (2005), MarkIT in Action - A Report on the Automatic Grading of DET Year 10 Essays, Curtin University, Perth, Australia.

I was the sole author of this paper. It was a non-refereed report commissioned by the Western Australian Department of Education and Training (WADET). It reported the findings of a trial of the MarkIT system with WADET and Year 10 student essays on the topic of the "School Leaving Age". An important question with the new system was whether it could process a large number of real student essays produced under formal testing conditions. The conclusion of the trial was that they could, and WADET subsequently became an industry partner in a successful Australian Research Council (ARC) Linkage Grant valued at \$740,000.

Paper 6.

Williams, R (2006), The Power of Normalised Word Vectors for Automatically Grading Essays, *Journal of Issues in Informing Science and Information Technology*, vol. 3, pp. 721-729.

I was the sole author of this refereed paper. The paper discussed the use of Latent Semantic Analysis (LSA) and Normalised Word Vectors (NWV) for grading essays. It also described further theoretical and practical work that was undertaken to test the author's ideas on an alternative way of grading essays.

Paper 7.

Williams, R (2007), A Computational Effective Document Semantic Representation, *Proceedings of IEEE-Digital Ecosystems and Technologies 2007 Conference*, Cairns, Australia, pp. 410-415.

I was the sole author of this refereed paper. The paper discussed a technique for representing the semantics of a document. After gaining an understanding of some of the existing AEG systems, and their features and ability to grade effectively real student essays, the author thought that it might be possible to approach the AEG problem in a novel way. This paper described the resultant theory and algorithms.

Acknowledgements

I wish to thank my professional colleagues Professor Elizabeth Chang (a member of my PhD Supplication Review Committee) and Professor Heinz Dreher, both of Curtin University, for their support and encouragement in the conduct of this project.

I also wish to thank the other members of my PhD Supplication Review Committee – Professor Leonie Rennie, Professor Graeme Wright, Professor Shelley Yeo and Professor Geoff West – for their valuable guidance to me for the completion of this exegesis.

The project could not have succeeded without the love and support of my wife Therese and my daughter Claire, both of whom endured many days and nights of discussion about the project, and shared the highs and lows of the software successes and failures. Thank you Therese and Claire!

Chapter 1 Motivation for the Study

1.1 Introduction

Automated Essay Grading (AEG) technology has been in demand by a number of education institutions, particularly in the United States, and it is envisioned that it will be increasingly deployed in countries around the world in the next 10 years. Not only can AEG systems provide substantial feedback beyond that which a teacher can provide when doing manual marking, they also can provide extensive quality feedback to both teachers and students. This helps improve the teaching and learning experiences by all parties involved.

In this Chapter, an introduction to the preliminary knowledge about automated essay grading is given, followed by the general problems associated with this technology that formed the motivation for this research. The significance of the research is then stated, followed by an outline of the structure of the exegesis.

1.2 Preliminary Knowledge about Automated Essay Grading

1.2.1 An essay

When teachers are evaluating students' learning, essays are preferred over multiple choice and true/false tests in many circumstances. Essays tend to allow students to demonstrate the acquisition of higher learning skills and the development of arguments for their point of view on the topic. However, an essay takes more effort and time for a teacher to grade than an objective test. An essay provides an indication of the nature and quality of students' thought processes, as well as their ability to argue in support of their conclusions (Ebel, 1979).

The essay defined by Cohen (2006) is as follows:

“Essay: The debate of issues surrounding a given topic. Essays use a formal structure and formal language to construct a persuasive argument. Skilfully done, the essay includes an introduction presenting the exegesis statement and background details; paragraphs, each of which present an idea to support the exegesis together with supporting evidence; and a conclusion which reflects on the implications of the exegesis. The MSE9 Essay marking guide assesses the narrative form through the criteria *Introduction, Argument and Ideas* and *Conclusion*.”

(Cohen, 2006).

In Australia each year students in years 3, 5, 7 and 9 across the country are tested for literacy. Part of this assessment encompasses the students writing a short narrative essay. About 1 million essays are marked each year. In Western Australia, students in Year 12 write essays in their final examinations. Many university undergraduate programs in business studies also require students to write essays.

1.2.2 Essay test

Essay testing generally involves the measuring of the higher-level abilities of students, rather than simple factual recall. Ebel comments on essay tests as follows:

“... [essays] provide a better indication of students’ real achievements in learning. Students are not given ready-made answers but must have command of an ample store of knowledge that enables them to relate facts and principles, to organize them into a coherent and logical progression, and then to do justice to these ideas in written expression.”

(Ebel, 1979, p. 96).

1.2.3 Essay grading

The grade assigned to an essay is an indication of the quality of the essay. Essay Grading is the process in which an essay is under an examination or assessment. Ebel states that

“An essay examination is relatively easy to prepare but rather tedious and difficult to score accurately. A good objective examination is relatively tedious and difficult to prepare but comparatively easy to score.”

(Ebel, 1979, p. 100).

A typical essay grading process involves reading the essay, understanding the content, checking grammar, looking at the cohesion of the story line, looking for correct syntactic and semantic wording of the sentences, and picking up errors in spelling.

1.2.4 Essay grading challenges

The grading of essays by teachers is very time consuming. Generally, teachers are given limited time allowances to grade essays. Payment for grading is generally based on the marking effort involved, but in many cases the time allowance is less than desirable, in order to limit costs. This means teachers are tempted to limit the number of essay assessments they give. This in turn affects feedback given to the students. Feedback is essential for practical improvement of essay writing. Students then do not get enough critiques and discussion of practical issues of their essay writing.

Feedback is important because students learn from corrections to their grammar, and the provision of alternative ways of expressing ideas.

1.2.5 Computer aided grading system

Computer based assessment began in 1955 (Lindquist, 1955). Lindquist developed optical test-scoring equipment at the University of Iowa. Large-scale testing programs, involving millions of students at all educational levels, are now commonplace. These programs are made efficient and effective with computer and scanning technology (Baker, 1976). This equipment however is only suitable for true-false and multiple choice questions, commonly known as objective tests.

Computer support for scoring objective tests is widely available. For example, mark sense sheets are used regularly to assess students enrolled in the undergraduate unit Business Information Systems 100 at Curtin University. Thousands of these sheets are processed each year. The challenge is to move from the simplistic character recognition to a more complex assessment of content as required by essays.

This exegesis describes an intelligent Automated Essay Grading System designed and developed by the author, which can grade essays efficiently, and help teachers and students at all levels. This help is provided in the form of an essay grade, feedback on spelling and grammatical errors, and interactive visual feedback on missing essay content.

1.2.6 Automated essay grading (AEG) systems

Research on AEG systems has been undertaken since 1966 when Page began his pioneering work (Page, 1966). His work was motivated by the need to reduce the high grading load that school English teachers faced.

An AEG system, sometimes referred to as an Automated Essay Scoring (AES) system, is a computer based system that analyses the content and/or grammatical style of electronic versions of textual essays in order to assign grades to the essays. Shermis and Burstein give the following definition:

“Automated essay scoring (AES) is the ability of computer technology to evaluate and score written prose.”

(Shermis & Burstein, 2003, p. xiii).

The aim of AEG systems is to perform at least as well as human graders, and preferably better. Performance can be measured in terms of agreement with the scores assigned by human graders, as well as the time it takes to perform the assessment. The AEG systems are generally more economical

and efficient when large volumes of essays on the same topic are to be graded; for example when at least several hundred, several thousand, or even hundreds of thousands of essays, are to be scored. For example, the Graduate Management Aptitude Test (GMAT) program has in excess of 600,000 essays graded by computer, and humans, each year. AEG systems can generally halve the cost of grading in comparison to using human graders when used for large numbers of essays. AEG systems can also be used in conjunction with human markers in order to detect discrepancies in the human grades.

1.3 Problems Associated with Automated Essay Grading Systems

1.3.1 Summative vs. formative assessment

AEG systems can be classified into two types: those that grade for style, and those that grade for content (Page, 1966). However, hybrid systems are starting to emerge which attempt to do both tasks. Teachers really require more from AEG systems than simply a grade for style or content if the systems are to be widely accepted. It is important that AEG systems provide useful feedback, both summative and formative. **Summative assessment** simply provides a grade summarizing the standard of the essay. **Formative assessment** provides information to the students so that they can improve the essay. Most existing systems provide summative assessment; a few include some formative assessment.

1.3.2 Non-interactive feedback

Another problem is the feedback is generally in the form of graphical indicators representing the levels obtained on various measures, or reports produced indicating deficiencies in the essay and improvements that could be made. The user cannot interact with these feedback items to explore improvements that could be made based on the feedback.

1.3.3 Limited computation power

Ellis Page has the honour of being the first published author on AEG systems. His seminal 1966 paper entitled "The Imminence of Grading Essays by Computer" (Page, 1966) contained the first attempt to outline a theory, and an implementation, of an AEG system. However, the inadequate mainframe computer technology delayed the research and development of a practical system for the real world of classroom essay grading. The work lapsed until the 1990s when the computer technology was more facilitating for the field (Page, 1994) and Page was the first founder of a successful and usable AEG system (Page, 1994; Page & Peterson, 1995). The advances in computer technology during the 1990-2000's, such as the increased power of Natural Language Processing (NLP), and the advances in

document processing and document retrieval, enabled today's researchers to get more involved in AEG research and development.

1.3.4 Fooling AEG systems

Recent criticism has been for the perceived ease with which some of the AEG systems can be fooled into scoring "nonsense" essays well; for example, some systems will assign the same score when the order of the sentences in an essay is reversed. McGee found this when he trialled the Intelligent Essay Assessor (McGee, 2006). Powers and colleagues reported that a professor of computational linguistics fooled E-rater into giving a top score of 6 for an essay that consisted of several paragraphs repeated 37 times, whereas the human graders awarded the lowest score of 1 (Powers, et al., 2001).

1.4 Motivation for the Study

There are four problem areas in the field of education that AEG systems could help address. These problems motivated the author to conduct this study in the field of AEG systems. These stimuli are:

- the technology adoption challenges in the education sector,
- the cost to Governments of administering literacy testing,
- the effort of marking essays by teachers,
- the value of feedback for students.

1.4.1 The technology challenges

Computer based assessment up to the 1990s was only suitable for true-false and multiple choice questions, commonly known as objective tests. However, computer based objective tests cannot

"... include such outcomes as the ability to recall, organize, and integrate ideas; the ability to express oneself in writing; and the ability to supply rather than merely identify interpretations and applications of data."

(Gronlund & Linn, 1990, p. 211).

As indicated in section 1.2.4, the problems with existing systems are mainly focused on the summative rather than the formative assessments, and the feedback is largely non-interactive. An example of summative assessment is simply the provision of a percentage mark for an essay e.g. 72%. Non-interactive feedback entails the provision of written comments on the essay document, which does not allow for interaction with the essay content, such as is possible with electronic documents and a word processor. For example, word processors allow searches for key word

occurrences in an electronic document. A teacher may wish to search for key words in order to see if a student has covered the required content.

1.4.2 The cost to governments

Primary and Secondary education systems are largely funded by governments, and have been for over one hundred years. Mason and Grove-Stephenson estimate that 30% of British teachers' time is spent on marking at a cost of 3 billion UK pounds per year (Mason & Grove-Stephenson, 2002). Computer support for scoring essays would lessen the extra time spent by teachers on marking essays in comparison with objective tests, and reduce the associated costs.

In Australia, Year 3, 5, 7 and 9 students undergo testing for literacy and numeracy under the National Assessment Program Literacy and Numeracy (NAPLAN) each year (NAPLAN, 2010a). About 1 million assessments are made each year. Prior to 2008, Western Australian Year 3, 5 and 7 students (about 90,000) participated in literacy and numeracy testing under the Western Australian Literacy and Numeracy (WALNA) program. Approximately 300 human graders were involved, all of whom underwent specialised training to ensure uniform grading. It cost WADET about \$500,000 to grade these essays. It is estimated that under NAPLAN, costs are about \$5,000,000. Substantial savings could be made if automated grading could be used for these essays.

1.4.3 The effort made by teachers

Scoring essays is time-consuming, and a large drain on human resources, especially on teachers' efforts. Teachers may be tempted to limit the number of essay assessments they set for their classes because of the effort involved in grading them. For disciplines where essays are an important form of assessment, this may limit the quality of assessment undertaken.

1.4.4 The value of feedback

Race (2005) indicates that feedback should enable students to understand the assessment task they have undertaken, help them own the need to learn, and motivate them to learn. Formative assessment is obviously more appropriate than summative assessment in meeting these requirements.

1.4.5 Technology adoption in the education sector

The education sector has been slow to adopt computer technology for teaching and learning and has not gained the productivity increases from computer technology which other industries have enjoyed (Ullman, 2007). The fees students pay for a college degree in the United States have increased

substantially in real terms in the past 40 years. Ullman (2007) provides the following data to show the lack of productivity gains in education compared with the United States telephone and mail systems:

Table 1.2. College Tuition Costs v Telephone Costs (Source: Ullman, 2007, Slide 3)

	Tuition	3-min LD call	Ratio
1959	US\$1,200	US\$3.00	400
2004	US\$30,000	US\$0.15	200,000

Table 1.2 shows that from 1959 to 2004 the annual cost of college tuition in the US increased from \$1,200 to \$30,000. It also shows that in the same period, a 3-minute long distance telephone call in the US decreased from \$3 to 15 cents. On this basis, the ratio between the two costs, originally 400, has increased to 200,000. This indicates that the efficiency in the education sector, using college tuition costs as a measure, has not kept pace with that of the telecommunications sector.

Table 1.3. College Tuition Costs v Postage Costs (Source: Ullman, 2007, Slide 5)

	Tuition	Airmail Stamp	Ratio
1959	US\$1,200	US\$0.08	15,000
2004	US\$30,000	US\$0.37	81,000

Table 1.3 shows that from 1959 to 2004 the annual cost of college tuition in the US increased from \$1,200 to \$30,000. It also shows that in the same period, the cost of a standard airmail stamp in the US increased from 8 cents to 37 cents. On this basis, the ratio between the two costs, originally 15,000, has increased to 81,000. This indicates that the efficiency in the education sector, using college tuition costs as a measure, has not kept pace with that of the mail system.

It can be argued however that the telephone system and the mail system are more suitable to automation through the use of IT than the education sector. Teachers' salaries are a major cost in the education sector, and continually increase, and it is difficult to replace the humans in the classrooms.

However AEG systems are one of the categories of information systems technology that can substantially contribute to increasing productivity in the education sector. This usage may result in similar productivity gains, which have been made in other sectors of the economy.

1.5 The Objectives of the Exegesis

The objectives of this exegesis are to present a proof of concept for an AEG technology, its significance for the education sector, including primary, secondary and tertiary education, and associated challenges and technical problems.

The aims are:

1. To give a comprehensive literature review of the current state of the art and emerging technology for the phenomenon of Automated Essay Grading systems.
2. To develop a theoretical foundation for the new and unique AEG system named MarkIT.
3. To design the architectural framework and to build a commercial grade prototype system, based on the theoretical foundation, and to prove the concept by the evaluation with a large scale performance test of MarkIT with the Australian Research Council (ARC) industry partner the Western Australian Department of Education and Training, and finally to present the results.

1.6 Significance of the Research

Teachers around the world spend considerable time preparing, administrating and marking assessments, many of which are essays. As noted earlier the assessment of essays involves a significant proportion of teachers' time. In many countries, governments are mandating standardised testing to monitor the effectiveness of their education systems. The Western Australian Department of Education and Training estimates that it costs the Department up to \$500,000 each year to mark the English literacy test essays for the Federal National Assessment Program – Literacy and Numeracy (NAPLAN) program. Much of the cost is incurred through training up to 300 markers and paying their wages.

Many educators, particularly in the tertiary sector, now make use of information technology to deliver their courses. Learning management systems (LMS), such as Blackboard, are used to post lecture notes, assignments, quizzes, marks and notices. These LMSs would be considerably enhanced if they could also automatically mark essay assignments.

AEG systems also allow for almost immediate feedback in contrast to the several days or weeks it takes to provide feedback with manual marking. Much of the mechanics of an essay, such as spelling and grammar, can also be extracted with AEG systems, freeing up the teacher to focus on higher level aspects of the essays.

AEG systems, if widely adopted, can therefore make a considerable positive impact on the workload of teachers in relation to assessments. Document processing techniques developed during the AEG development process will also have applications to other areas of text assessment.

1.7 Overview of the Exegesis

This exegesis consists of eight chapters.

Chapter 2 discusses the literature on existing AEG systems. The theoretical underpinnings of the different systems are described and their key features summarized. The emphasis placed on grading for style or content for each system is also discussed. Finally a table summarising the performance of each system is given.

Chapter 3 discusses the problem definition for the new AEG system known as MarkIT. It discusses the shortcomings in several of the existing systems, and how MarkIT attempts to address some of these shortcomings. The weaknesses in the theories underpinning the existing systems are highlighted, and the reasons for developing a new theory are given.

Chapter 4 provides an overview of the theoretical foundations of the proposed solution for the AEG system, known as MarkIT.

Chapter 5 discusses the design and implementation of the MarkIT system including its architecture and feedback mechanisms.

Chapter 6 describes the testing and evaluation of the system using 390 essays written by Year 10 high school students on the topic of the School Leaving Age.

Chapter 7 consists of published scientific work on the implementation of the theoretical foundation in the design and development of a commercial grading system. It also discusses the testing of the system with a large volume of essays from WADET and a discussion of the results.

Chapter 8 summaries the exegesis, the limitations of the work, and the future work which could be undertaken to improve the system.

1.8 Summary

Essays are preferred over objective tests when teachers wish to assess their students' high level learning skills, and the students' ability to develop arguments to persuade the reader to their point of view. However, essays require considerable more time and effort to grade than objective tests. The provision of formative feedback for essays is very time consuming, particularly when there are large numbers of essays to be graded. Computers have been used for many years to grade objective tests. Automated essay grading systems are starting to emerge into mainstream education to help overcome the manual grading time and costs. This exegesis reviews the literature on AEG systems, develops a new theoretical framework for an AEG, describes the development of a new system named MarkIT, and discusses the results of a trial of MarkIT.

In the next chapter an introduction and discussion of the field of AEG systems is given. In particular, the problems associated with AEG systems are discussed, which has provided the motivation for this research and defined its objectives. The significance of the research and the structure of the methodology are discussed. The next chapter also presents the comprehensive literature on AEG systems and gives a critical evaluation of the strength and weaknesses of the existing AEG systems.

Chapter 2 Literature Review

2.1 Introduction

This chapter reviews the literature on AEG systems. It discusses the features and performance of a number of systems, some of which are available commercially. Their theoretical foundations are highlighted, as well as the techniques used, the performance, and a comparison of the overall advantages and disadvantages of each system. A summary table is then presented comparing and contrasting each of the systems.

2.2 Preliminary Concepts

2.2.1 The “Style” in the essay

Essay style refers to the argumentative and grammatical structure of the essay. Good style includes correct grammar, logical consistency, and conformity to the accepted structure for essays in the targeted subject domain.

2.2.2 The “Content” in the essay

Essay content refers to the subject matter discussed in the essay. High content marks are given to essays that include good discussions of all the necessary material sought by domain experts.

2.2.3 The “Systems”

The “systems” in this exegesis refer to the commercial grade AEG systems that are in use, or under development, for use by educational institutions. These systems have unique names, authors and developers. Many of these systems primarily mark for either “Content” or “Style”, but sometimes for both. They are based on different scientific foundations such as Natural Language Processing, statistical techniques and neural networks.

2.2.4 The scientific approaches that underpin the “Systems”

Natural Language Processing (NLP) involves the use of computer technology to process natural language in order to understand its meaning. Typically, it involves the use of parsing algorithms to construct semantic structures, which are then used to determine meaning.

Statistical NLP is becoming pervasive: these techniques use probability and corpus data to detect patterns in texts and to derive meaning from them. Some AEG systems also make use of multiple linear regression techniques.

Other scientific techniques include Bayesian, Neural Network, and Fuzzy methods.

2.2.5 Rubric

A rubric is a marking guide prepared to enable markers to assess an essay. Generally, a rubric contains details about the structure, quality and the content expected in an essay, and a breakdown of the marks distribution allocated by the rubric to the components of the essay.

2.3 Overview of Scientific Approaches from the Literature

Today there are a variety of active communities researching the topics associated with automated essay grading. Thirteen AEG systems have been identified. Details of some of these systems were first compared and contrasted by Williams (2001), followed by Valenti, et al. (2003), and Lam, et al. (2010). These systems include:

1. Project Essay Grade,
2. Intelligent Essay Assessor,
3. Educational Testing Service 1,
4. Electronic Essay Rater,
5. Conceptual Rater,
6. Bayesian Essay Test Scoring System,
7. Intelligent Essay Marking System,
8. AutoMark,
9. Schema Extract Analyse and Report,
10. Paperless School Free Text Marking Engine,
11. Text Categorisation Technique,,
12. Intellimetric,
13. SAGrader.

The above systems are differentiated by the use of one, or several of, the following underpinning scientific approaches:

- Statistical based approaches,
- LSA (Latent Semantic Analysis) based approaches,
- NLP (Natural Language Processing) based approaches,
- Bayesian based approaches,
- Neural Network based approaches,
- Semantic based approaches.

In many cases, NLP is used in conjunction with the underlying technique. In the next sections, a detailed survey of all the scientific approaches that underpin the existing AEG systems is given.

2.4 Statistical Based Approaches

Multiple linear regression is the predominant statistical technique used. Generally, many essay features are extracted from training essays using NLP, and these features are used as independent variables to build an equation to predict the dependent variable, which is the essay score. The literature discussed two systems primarily based on statistical techniques. These were

1. Project Essay Grade (PEG),
2. Paperless School Free Text Marking Engine (PSFTME).

2.4.1 Project Essay Grade (PEG)

PEG was developed by Page (1966), and was the earliest reported AEG system in existence. The system primarily graded for linguistic features, rather than essay content. A regression equation was computed for these features, as measured in training essays, and then applied to the essays to be graded. In the early days of development, Page found it difficult to obtain, through computer analysis, the linguistic variables he wanted to use in the regression analysis.

Page focused his research on the use of surface features of essays, such as the number of words in an essay and other linguistic features, to predict essay scores. In one published regression analysis thirty four predictors were used – not all of them however were statistically significant. Page reported many human-computer score correlations, the highest being 0.88. However, he also recommended that further research be undertaken to be able to identify predictors that would identify essays that were graded significantly higher by the system than by the humans.

PEG achieved some high correlations with the grades it assigned and the grades given by human markers. However, it had a number of weaknesses. The measures of essay quality that the system used were not intrinsic measures, but indirect approximations, such the number of words and the number of propositions (Hearst, 2000). It also did not assess the content, style and organisation of the essay, and feedback on these important criteria was not provided (Hearst, 2000).

Page set the groundwork for development of AEG systems, and many of the newer systems make use of some of his ideas.

2.4.2 Paperless School Free Text Marking Engine

The Paperless School Free Text Marking Engine was developed by Mason & Grove-Stephenson for the scoring of low stakes essays and short tests (Mason & Grove-Stephenson, 2002). The authors' motivation to build the system was influenced by the facts that teachers in the UK spent 40% of their time in the classroom, and another 30% in marking. Automated aids to support marking would lead to an increase in teachers' productivity, as less time would be spent on marking, and the extra time would be applied to higher level teaching tasks. The system made use of NLP techniques to analyse grammatical aspects of the essay, extraction of meaning, and finally determination of the relevancy of the extracted information to a correct answer. The theoretical basis of the system was Bloom's taxonomy, comprising the six levels of cognitive skill of knowledge, comprehension, application, analysis, synthesis and evaluation (Bloom, 1956). The specific subsets used by the system were

- Knowledge – evaluated against a list of concepts derived from, for example, textbooks on the topic,
- Understanding – details were not available as the algorithm was commercially sensitive and not reported,
- Evaluation – a refinement of counting of adjectives and adverbs.

The system was incorporated in a wider Learning Management System (LMS). Student essays were submitted to the LMS server, appropriate master texts identified (both positive and negative), scoring weights determined when student essays were compared against the master texts through regression analysis, grades assigned, and automatic selection of appropriate comments undertaken. This enabled both summative and formative assessments to be given to the students. The authors found that real time scoring was not possible, and so the system was implemented as a queuing system. Problems identified by the authors were:

- human to human agreement on scores was variable, affecting calibration,
- difficulty in selection of appropriate master texts,
- grading of graduate level essays,
- grammatical and spelling errors.

2.5 Latent Semantic Analysis Based Approaches

Latent Semantic Analysis is based on the work of Salton (1989) in Latent Semantic Indexing (LSI) and text retrieval systems. In LSI a relatively small number (e.g. 5 – 10) of significant document terms that determine the document content are established. Each of these terms is regarded as a dimension in a vector space. The content of a document can then be represented by the vector formed by these

terms. LSI theory assumes that similar documents will have similar term vectors. To retrieve documents from a collection a search term vector is constructed representing the desired content and then documents with similar vectors are retrieved.

2.5.1 Latent Semantic Analysis (LSA)

Latent Semantic Analysis uses a derivation of LSI. A semantic space is built from a matrix constructed for all the words in a training collection of graded essays (or content documents). Each unique word occurring in the documents is allocated a row in a two dimensional matrix. Each document is allocated a separate column in the matrix. The number of times each word occurs in each document is placed in the cells at the intersections of the rows and columns. This matrix is then processed by a linear algebra technique known as Singular Value Decomposition, which reduces the dimensions of the matrix. Term vectors are then constructed for the graded essays in this space. A term vector for an ungraded essay is built and matched to the closest scored vectors in the semantic space, and then assigned the corresponding grade. A system based on LSA is the Intelligent Essay Assessor (IEA).

2.5.2 Intelligent Essay Assessor (IEA)

The IEA system was developed by Landauer and his colleagues (Landauer, et al., 1998). The system at the time primarily graded for content rather than literary style. It was based on LSA.

In practice, the semantic space was built from electronic texts on the topic to be graded, or from several hundred human graded essays. Typically, hundreds of vectors representing graded essays were represented in this space. An un-graded essay was then processed to produce a vector in the semantic space, and the essay was given the grade of the nearby graded essay vectors. Closeness was measured by the cosine between the two vectors. The IEA produced agreement between human and computer graders as high as 91%. Landauer et al. (1998) suggested further work needed to be done by analysing larger written and spoken texts to see if the LSA approach was still effective. Landauer's work produced an alternative way of processing essays for grading, which led to other researchers using a similar approach. The requirement for at least one hundred training essays on a particular topic means that the IEA is not suitable for assessing essays from small classes.

2.6 Natural Language Processing Based Approaches

NLP has been a topic of research for over 50 years. Intuitively it was thought that NLP could be applied productively to essay grading, and by and large, systems based on NLP have been successful. It is the most common paradigm for AEG systems.

2.6.1 Natural Language Processing (NLP)

NLP makes use of language grammar and syntax structures by computers to automatically understand and use natural languages, such as English, French and German.

Six AEG systems that utilise the NLP approaches were found in the literature. NLP-based systems are especially suitable for marking at the content level, not just for style. These systems are:

1. Educational Testing Service 1 (ETS 1),
2. Electronic Essay Rater (E-Rater),
3. Conceptual Rater (C-Rater),
4. AutoMark,
5. Schema Extract Analyse and Report (SEAR),
6. Intellimetric.

In the following sections, each system is discussed.

2.6.2 Educational Testing Service 1 (ETS 1)

The ETS 1 system was developed by Burstein and her colleagues (Burstein, et al., 1996). The developers wanted to build a system that would grade short-answer free-responses to test questions. The core of the system was a domain-specific, concept-based lexicon. Concept grammar rules were manually built from training set responses for this purpose. It was estimated that forty hours of manual work were required for this task. The system looked for similarities in meaning for multiple responses, and discounted similar responses when computing a total score for a student's responses to the prompt. Subsequently about 90% accuracy was obtained from the validation responses.

This system was limited to grading short answers, and so was not useful for grading longer essays. The major weakness of the system however was the requirement to build the lexicon manually for each knowledge domain. Such a requirement meant that this time consuming task was onerous and hence the system would not gain widespread acceptance by teachers.

2.6.3 Electronic Essay Rater (E-Rater)

The E-Rater system was also developed by Burstein and her colleagues (Burstein, et al., 1998). In comparison with ETS 1, this system attempted to grade essays written by students in answer to prompts. A prompt is the term used to describe the essay topic. Statistical and NLP techniques were used to discover linguistic features of the essays. The system was trained on a sample of human graded essays on the topic in question, and a scoring algorithm built on patterns of features, which

were significant in a multiple regression analysis. It graded for writing style and essay content. Typical predictor features of an essay score were items such as

- Argument content score,
- Essay word frequency content score,
- Total argument development words/phrases,
- Total pronouns beginning arguments,
- Total complement clauses beginning arguments,
- Total summary words beginning arguments,
- Total detail words beginning arguments,
- Total rhetorical words developing arguments,
- Subjunctive modal verbs.

E-Rater performed well in assessing syntactic variety, well developed arguments, and essay content. In its current form, it also provides feedback to students about their essays. The requirement for a large number of training essays meant that it was not suitable for assessment for small classes.

2.6.4 Conceptual Rater (C-Rater)

The C-Rater system was related to E-rater in that it shared common technology and was developed by Burstein and colleagues (Burstein, et al., 2001). However, its purpose was different. It aimed to score short essays for content written in response to prompts such as those that were found at the end of text book chapters as short exercises. It also did not need training on hundreds of human graded essays, but simply needed a single correct answer to perform the grading task. While E-rater produced a holistic score for the student's skill in writing, and did not score for specific content, C-rater could distinguish between a correct and incorrect answer. It did this by detecting whether specific concept information was present in the student response. C-rater had the advantage over some other AEG systems in that it did not require training on a collection of human graded essays. It was also relatively easy to setup for use. However, it was not suitable for grading longer essays.

2.6.5 AutoMark

AutoMark was developed by Mitchell and colleagues to mark short free text responses for specific content (Mitchell, et al., 2002). It evaluated spelling, typing, syntax and semantics in the responses. This content was specified in a number of manually prepared marking scheme templates that structured the acceptable and unacceptable answers into simple sentence structures. An example of a structure the authors provided was one consisting of nouns, verbs and prepositions. Information extraction was then performed on the student responses. Information extraction made use of NLP

techniques, but instead of performing in-depth language analysis, it simply looked at the surface level aspects of a sentence to extract specific concepts. The authors tested the system on 11 year old UK students who were assessed as part of the national curriculum science assessment. Four types of answers for 120 students were used for the testing. These answers consisted of single word, single value, and a short explanatory sentence, all scored as 1 for a correct answer, and a data pattern description, scored out of 2. Overall performance was 93.3% accuracy for computer versus human scores. When revised templates were developed, based on material discovered in the first round of scoring, the performance increased to 96.5%. Limitations of the system included problems with incorrectly used words, and misspelled words. Problems also arose with an inability to analyse correctly sentence structures, because of poor structures in answers. Answers that contained an incorrect qualification were also problematic. Finally, complex answer structures could not be adequately expressed within the answer template structures. The system is now offered commercially, and details of this deployment can be found in Intelligent Assessment Technologies Ltd (2006).

2.6.6 Schema Extract Analyse and Report (SEAR)

SEAR was developed by Christie to grade for both style and content (Christie, 1999; 2003). The system had an extract phase whereby essays in Microsoft Word® format were taken from a submission file folder and placed in an extract folder for processing. The essays were then marked for style, content, or both. Separate software was used for each process. To assess content, a structured content schema was developed for each question. This schema was relatively simple to set up; topics expected to be covered in a correct answer were listed, together with qualifying material. Marks for each section were also allocated. The student answer was then compared to the marking schema, and marks allocated based on the amount of schema content covered by the student. Christie indicated that he had not field tested the style assessment function at the date of publication. The best human – computer scores correlation obtained in the testing was 0.60 compared to best human – human correlation of 0.81. These results were obtained from three different sets of essays that were oriented towards answers containing facts or knowledge, or an explanation of a process. A limitation of the approach embodied in SEAR was that the larger the content schema, and the marks allocated, the poorer the system performance. Feedback to students from the system was limited to a statement of facts not covered by the answer. Christie also discussed future improvements to his algorithms, such as providing information on plagiarism.

2.6.7 Intellimetric

Intellimetric was available commercially from Vantage Learning (Vantage Learning, 2004; 2005). It was based on the following key principles relating to current thinking about the human brain and its processes:

1. Modelled on human brain processes and how humans scored text documents,
2. Used a learning engine – evaluated essays based on how humans have scored essays,
3. Systemic – made use of many system components to make judgements,
4. Inductive – judgements were made from information obtained in a “bottom up” approach,
5. Multiple judgements – many types of information were processed by mathematical techniques.

It assessed an essay by considering over 400 essay semantic, syntactic and discourse features, which were grouped under the two categories of content and structure. Content features included, amongst others, breadth, consistency, cohesiveness, and relationships among the various parts of the essay, and idea sequence. Structure features included, amongst others, conformance to grammar, sentence complexity and completion, and syntactic variety. The system required between 50 and 300 human graded essays to train the system on a particular essay topic. The training essays were processed to extract the underlying features that humans were using to evaluate essays at each score point – typically of 3, 4 or 6 levels. This information was incorporated in mathematical processes to build system “judges”. Better performance was obtained if two or more humans graded each of the training essays. The essay grading process involved several phases including text pre-processing, text parsing, computational analysis, production of several judging scores, and provision of the final score. The system had access to a 500,000 word vocabulary and a 16 million word concept net, which was used at the computational phase. The system graded for focus and meaning, organisation, content and development, language use and style, and mechanics and convention. The algorithms used were not published in the literature as they were commercially sensitive – however, Vantage Learning (2005) acknowledged that regression analysis played a part in the processes. Intellimetric claimed that the system scores agreed with the experts’ scores about 95 to 100% of the time. A 6-point scale was preferred over smaller scoring scales.

Intellimetric is possibly one of the most substantial essay grading systems, in terms of its sophisticated algorithms.

2.7 Bayesian Based Approaches

These approaches make use of Bayes' conditional probability theorem to determine the likelihood that an essay belongs to a given (score) category. Two well-known systems are based on the Bayesian methods, namely:

1. Bayesian Essay Test Scoring System (BETSY),
2. Text Categorisation Technique (TCT).

2.7.1 Bayesian Essay Test Scoring System (BETSY)

BETSY was developed by Rudner and Liang and used Bayesian statistical methods to grade essays for content (Rudner & Liang, 2002). Bayesian statistical methods are based on Bayes' theorem, which provides a formula to determine the probability of an event occurring, given that another event has occurred. The system aimed to classify essays into four categories, namely extensive, essential, partial and unsatisfactory when measured against the essay requirements. However, in the trial of the system these four categories were collapsed into two categories; the counts of the system calibration essays in the top and bottom categories were extremely small. The system looked for probabilities associated with occurrences in essays of items such as specific words and phrases, presence of specific noun-verb pairs, and the order in which specific concepts were encountered. In the research presented in Rudner & Liang (2002), over four hundred and sixty essays were used to calibrate the system. The system correctly categorised eighty percent of the essays. While interesting, this system did not provide feedback. The coarse granularity, represented by only having two categories of classification, limited the usefulness of the system for many essay assessments. The relatively large number of calibrating essays required also limited the systems usage to classes with a large enrolment.

2.7.2 Text Categorisation Technique (TCT)

The TCT system was developed by Larkey and graded for content and/or style (Larkey, 1998). An experiment was reported by Larkey in which essays on the topics of social studies, physics and law were used to train the system, and then test sets of essays graded by the resultant trained systems. Several different techniques were used.

Firstly, the training essays were processed by Bayesian classifiers. These could distinguish good essays from bad essays, rank them, and assign grades to them. They estimated the probability of a document being similar to an exemplar document based on the occurrence of certain words in the

document. Features to be used in the classifiers, based on stemmed words, were then chosen on the basis of training essay scores that had the highest correlation with the human scores.

Secondly, a k-nearest neighbour classifier was used. These classifiers looked for the “k” training set essays most similar to the test essay. The test essay was then assigned a grade, which was a similarity-weighted average of the human assigned grades of the k retrieved training essays.

Thirdly, eleven text-complexity features were identified, such as the number of words in a document, the number of sentences in a document, and average word length, amongst others. Stepwise linear regression was then performed on these features to find a suitable scoring equation.

Fourthly, regression analysis was performed on the Bayesian classifiers’ variables, and a scoring equation found.

Finally, all the variables from the three approaches – 11 text-complexity variables, k-nearest-neighbour scores, and the Bayesian classifiers’ scores- were used in linear regression to find a suitable scoring equation.

Human-computer correlations from these approaches ranged from 0.69 to 0.78 for the social studies essays, and 0.53 to 0.63 for the physics essays.

Another experiment was conducted on essay sets on general questions for college students wishing to undertake graduate studies. Correlations ranged from 0.69 to 0.88.

2.8 Neural Networks Based Approaches

Neural network approaches make use of artificial neural networks to learn about the characteristics of essays and then be able to assign scores based on these characteristics and their patterns.

2.8.1 Neural networks

Neural networks make use of non-linear statistical data modelling techniques to simulate connections between artificial neurons. These networks can detect complex relationships between data inputs and outputs. The networks are trained on sample data, and once a model is established, they can process unseen data, which in the context of this research, would be essays with the aim of assigning scores. One system based on a neural network was found in the literature, the Intelligent Essay Marking System (IEMS)

2.8.2 Intelligent Essay Marking System (IEMS)

The IEMS was developed by Ming and colleagues at the Ngee Ann Polytechnic in Singapore. The system's focus was on grading short essays for content. The technology was based on an Indextron, which was implemented with a Pattern Indexing Neural Network. (Ming, et al., 2000). An Indextron was described as a specific clustering algorithm that was implemented by the neural network. The Indextron can be explained as follows:

If there is a function $y = f(x)$, then generally there will be many x values for each y .

We can define a function index $fi(y)$ as the set of all the x values that map to y .

Ming et al. (2000) give the following example:

“Rating index $ri(y)$. Let all chess players x make up the domain of a function $r(x)$, whereas the players' ratings y make up the range of $r(x)$. For each rating value y , the rating index $ri(y)$ extracts a subset $\{x\}$ of players that have the same rating y .”

(Ming et al., 2000, p. 4)

Eighty five students used computers to write essays on the topic of crime in cyberspace. These essays were then graded by humans before the training process. The neural network was then invoked and it looked for essay feature patterns and used the human grades to build the function index. This index related essay feature vectors to the human grades. The correlation between the human and Indextron scores was initially 0.75, and was improved to 0.80 after a revised marking schema was used for a second round of training. The essays were marked in 1 – 2 seconds and the students could obtain immediate feedback on their essays.

2.9 Semantic Based Approaches

Systems based on semantics are targeted at assessing the actual content of an essay to assign a grade. Essays with more relevant content are scored higher than those with less. Typically, these systems must have a way of storing knowledge about the essay topic, and use various semantic structures to do this.

2.9.1 Semantic network

A semantic network is a structure containing words or concepts, and connections between these concepts or words indicating hierarchy and similarities in meanings. These networks can be navigated to find details of the relationships between target concepts or words. A literature review found one system based on this approach, namely SAGrader.

2.9.2 SAGrader

SAGrader was a commercially available system, which graded essays mainly for content (SAGrader, 2010). The system accessed semantic networks containing the specific knowledge for individual essay topics. The networks contained concepts, and their relationships, as well as features which could be used to find concepts in the student essays. These networks were built manually for each subject area. The system was suited to essay topics where there were a finite number of possible answers. It was not recommended for assessing creative writing essays.

After students submitted their essay to a web site, the essay was graded within seconds and appropriate feedback provided.

The feedback provided included the learning objectives which were correctly covered by the student, and those that were not covered. The system could be used in such a way that a student would submit a revised essay, after modifying an original essay in line with feedback received from a previous submission. Some students would submit up to six revisions before finally submitting their work. In one study, an 11% improvement in scores was obtained when this process was used.

The system provider studied student reactions to their use of the system, and found students liked the detailed and immediate feedback. They particularly liked the opportunity to submit revised essays. Many students also preferred automated essay grading usage for essay grading over sitting multiple choice tests. Students did not like the system when they incorrectly quoted material and were marked as incorrect for that particular part of the essay. They also did not like it when they were penalised because they quoted material that was not in the semantic network. In addition, they did not like it when they had the correct concept, but misspelled part of the idea under discussion. Some problems encountered by the system included correct material not in the semantic network, and the inability to process correctly material that was expressed differently to that in the semantic network.

Evaluations of four of these systems, namely PEG, E-rater, LSA and TCT, were published in the paper Williams, R (2001), *Automated Essay Grading: An Evaluation of Four Conceptual Models*, in Kulski, M & Herrmann, A (eds.), *New Horizons in University Teaching and Learning: Responding to Change*, Curtin University, Perth, Australia.

2.10 Phrase Processing and Term Vector Technologies

An important issue in essay grading is how to provide structures to represent meaning and understanding in essays. There are two key techniques commonly used for automated content understanding, namely: phrase processing and the term vector. These details are explained below.

2.10.1 Phrase processing

In order to establish some meaning from sentences in documents it is necessary to process at the level of individual phrases in the sentences. Typically, parsers do this in most NLP applications, and these processes are well understood. Processing is such that individual noun phrases and verb phrases are derived, and this is usually sufficient for the sentence semantics to be derived.

2.10.2 Salton's phrase processing theory

Salton developed a system for automatic phrase matching (Salton, 1966). He described phrase constructs and the use of thesaurus concept numbers for representing the content of documents. The particular application for which the structures were used was phrase matching for processing written texts. Salton describes best the situation at the time in relation to text processing when he refers to Luhn's suggestion of using the frequency of words in a document as a measure of the content (Luhn, 1957).

“More recently, the original statistical methods have been modified in various ways: by using word stems rather than the original word forms to identify document content; by introducing synonym dictionaries to lessen the effects of vocabulary variations; and, most importantly, by identifying relations between certain words to be used as content identifiers in conjunction with the surrounding words.

As a result, many of the word matching systems are now being replaced by *phrase* processing systems, in which the basic units being manipulated are sets of normalized words together with specified relations between them.”

(Salton, 1966, p. 169)

2.10.3 NLP phrase processing

Automated Essay Grading systems that make use of NLP generally parse sentences in a document into phrase structure trees for further processing. These structures are based on the established grammar rules for the language being processed. The most common phrases for the English language are noun phrases and verb phrases, and their structures are well understood. Consider the following sentence

“We are having hamburgers for dinner tonight.”

A possible parse tree for this sentence is shown in figure 2.1. Table 2.1 explains the mnemonics used in figure 2.1.

Table 2.1 Parse tree mnemonics

Mnemonic	Meaning
S	Sentence
NP	Noun phrase
VP	Verb phrase
N	Noun
V	Verb
P	Preposition

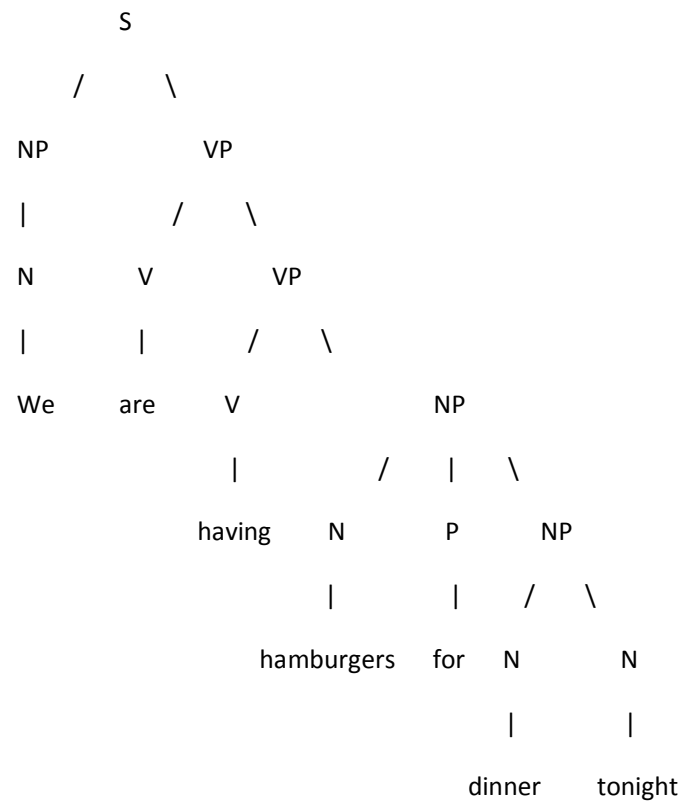


Figure 2.1 Sample parse tree

2.10.4 Term vector

To build a term vector for a document each unique word in the document forms a dimension of the vector. The counts of the occurrences of each word are then assigned as the index for that word's dimension. Typically, hundreds of dimensions are derived for a document. These vectors thus form a vector space, and similarities between different documents are calculated as the cosine of the angles between the vectors. Documents that are identical will have the same vectors, and hence be collinear, and have an angle of zero degrees between them. The cosine of this angle is therefore 1. Documents that are completely different will have vectors that are orthogonal, and an angle of 90 degrees between them. The cosine of this angle will be 0.

2.10.5 Term vector theory

This section reviews the literature on these vectors. Salton developed the document term vector representation theory for text retrieval applications (Salton, 1967; 1968; 1975a; 1975b; 1989, Salton & McGill, 1983). Terms in a document were allocated to the dimensions in a semantic space. In general, there were “n” terms in a document, and therefore “n” dimensions in the semantic space. A measure was assigned to each dimension – generally, it was the count of how many times that term occurred in the document. Weightings may have been applied to the terms as well – for example, the weights may have been assigned according to the relevant importance of the term in the context of the topic of the documents being retrieved. Salton also alluded to the use of a thesaurus for adjusting the terms.

“A thesaurus can be used to broaden the existing indexing vocabulary by replacing the initial terms with the corresponding thesaurus class identifiers, or by adding the thesaurus class identifiers to the original terms.”

(Salton & McGill, 1983, p. 81).

Comparisons between the content of documents could then be made by computing the cosine of the angle between two documents' vectors – the more similar the documents, the closer they would be aligned, and the closer to one would be the cosine (cosine takes a value between 0 and 1).

Siemens AG were granted a patent on a simple version of a thesaurus term vector representation (Siemens, 2000). This patent described a process which enabled the similarity between electronic documents to be examined by comparing term vectors which represented the content of the documents.

A thesaurus for the medical domain was constructed with Generic Terms. Each Generic Term had one or many Subordinate Terms. For example, the generic term “AIDS” had at least three

subordinate terms: HIV-1, HIV-2, and SIV. Similarly, the generic term “Pathogen” had at least four subordinate terms Virus, Bacteria, Microbe and Bacillus, etc. Term vectors were constructed. To construct these vectors, all words which were not contained as generic and/or subordinate terms in the electronic thesaurus were deleted in every electronic data set. The structure between the electronic data sets was determined from the generic term vectors.

“A development of the invention provides that each generic term vector has the number of a generic term contained in the respective electronic data set. In this manner the frequency of the generic terms occurring in an electronic document is also taken into consideration within the scope of the structure determination, leading to an even more reliable result of the comparison.”

(Siemens, 2000, p. 3).

2.11 Comparison of Scientific Methods Used

A summary of the 13 systems identified and their characteristics is shown in Table 2.2. This table presents the key features of the AEG systems identified in this research. The chief scientific method upon which they are based, and the main focus of the systems on either content or style, are identified. The identified scientific method is the primary method on which the system is based, but many of the systems make use of supporting techniques as well. It is common to see NLP used in conjunction with statistical methods. An AEG system’s main focus is on understanding an essay and assessing its relevance to the essay prompt. An essay prompt is the term used for the essay question in the essay grading community. As a consequence, NLP was seen as the obvious technique to use, and this is reflected in the fact that six of the systems use this as their primary technology. Some of these systems were developed to assess writing style in the context of detecting an essay writer’s level of competence. Others systems were developed to detect an essay writer’s understanding of a topic. The table summarises these emphases.

With regard to Phrase Processing, Salton (1966) extracted generic relationships between properties for phrases by:

- Performing a stem-suffix cut-off operation,
- Temporarily removing high frequency function words,
- Consulting a hierarchical arrangement of concept numbers,
- Replacing the word stems with thesaurus concept numbers.

With this processing, he was able to match phrases when applying his technique to a document retrieval application. However, he did not specify the structure of his phrases.

With regard to the use of Term Vectors, document retrieval techniques generally built vectors using only key terms in a document, and not all the words in the document. With LSA Landauer et al. (1998) use all terms in a document initially, but through the use of singular value decomposition ends up using about 400 terms in the vectors.

Table2.2. Comparison of AEG Systems

AEG System	Developer	Scientific Method	Main Focus
PEG	Page (1966)	Statistical	Style
PSFTME	Mason & Grove-Stephenson (2002)	Statistical	Content
IEA	Landauer et al. (1998)	LSA	Content
ETS 1	Burstein et al. (1996)	NLP	Content
E-Rater	Burstein et al. (1998)	NLP	Style and Content
C-Rater	Burstein et al. (2001)	NLP	Content
AutoMark	Mitchell et al. (2002)	NLP	Content
SEAR	Christie (1999)	NLP	Style and Content
Intellimetric	Vantage Learning	NLP	Style and Content
BETSY	Rudner & Liang (2002)	Bayesian	Style and Content
TCT	Larkey (1998)	Bayesian	Style and Content
IEMS	Ming et al. (2000)	Neural Network	Content
SAGrader	Idea Works (2010)	Semantic Network	Content

2.12 A Critical Review of the Existing Approaches to Building AEG Systems.

All of the above approaches and their systems perform reasonably well in achieving the goals set before development, namely grading for content and/or style. However, analysis found that they failed in the following key aspects, namely:

1. Many of them tended to derive scores from essay features, and not necessarily from a true understanding of the essay structure and content.
2. Most AEG systems require either training on hundreds of human graded essays on the specific topic, or manual construction of a knowledge repository. When the number of essays

to be graded is small, i.e. less than a few hundred, then this training task is not economical to undertake. Because of this, AEG systems are really only suitable when hundreds, or thousands of essays are to be processed.

3. AEG systems that use multiple linear regression techniques for building a scoring equation assume that there is a linear relationship between the essay features and a score. This is not necessarily true. It may be the case that assuming non-linearity leads to better outcomes, which of course requires different mathematical techniques.

There were four main sources of error in these AEG approaches identified by Valenti et al. (2003). These were:

- (a) An inability to identify correctly misspelled and misused words,
- (b) A failure to analyse properly the sentence structures,
- (c) An inability to identify as incorrect a qualification,
- (d) An inability to provide a mark scheme template.

This exegesis will address problems (a), (b) and (d) above. Problem (a) is addressed by using Microsoft Word to extract spelling and grammar errors in the essays. Problem (b) is addressed through the use of parsing sentences into pre-defined phrase structures. Problem (d) is addressed by the use of model answers that indicate the content that is expected.

Another challenge for the use of these AEG systems was to build trust amongst educators who needed to be convinced that the scores produced were valid, and that several humans would also have assigned the same score. The AEG system discussed in this exegesis addresses this issue by suggesting that clients provide grades from several humans for each training essay.

2.13 Summary

This chapter has presented a review of the current state of the underlying theories which existing AEG systems are based on. It has demonstrated that there are many theoretical constructs on which AEG systems are based. A detailed comparison of the existing systems' features was also presented, including the important question as to whether they grade for style or content. Some existing work on phrase structure representation and document term representation was also discussed. In many cases, the systems discussed closely matched the performance of human graders (Valenti et al., 2003).

In general, it can be concluded that there is not a single standard method of automatically grading essays – many approaches are successful. Some of these systems grade for content, others for linguistic style, and others for both. NLP algorithms are by far the most common techniques used. Statistical and LSI/LSA are the other most common techniques. All systems require training before use for a particular essay prompt. Generally between 50 and 400 human graded essays are required for training – thus restricting the use of AEG to high volumes of student essays.

Phrase structures have been experimented with since the 1950s in order to understand English sentences. The system described in this exegesis adds to this work by developing a new representation for phrases which is computationally effective i.e. a 400 word essay can be graded in 2-3 seconds.

At the end of this chapter, the author provides a comparison of the systems including the scientific methods that underpin the systems. This is followed by an evaluation of the systems discussed in the literature to date. The next chapter, chapter 3, builds on this knowledge, and the author defines the research problem addressed by this exegesis, and discusses the research methodologies that are considered to be appropriate.

Chapter 3 Problem Definition

3.1 Introduction

This chapter discusses the characteristics of an essay and state and national standard criteria used for grading essays, and technical issues associated with measuring these criteria. A discussion of the issues surrounding Information Systems (IS) research is then provided. The research method for the research discussed in this exegesis is presented and discussed.

3.2 Concepts and Definitions

3.2.1 The features of an essay

According to the definition of an essay in section 1.2.1, we see that the following aspects of an essay should be considered to determine a grade:

- Style – is the writing style appropriate for the type of essay requested e.g. is it a piece of academic work, or a fictional piece meant to entertain?
- Content – is the necessary and expected material covered sufficiently?
- Structure and organization – does it have an introduction, and a set of ideas argued for in a satisfactory way?
- Mechanistic aspects – is the spelling and grammar correct, and is there correct use of paragraphs?

3.2.2 National literacy and numeracy testing

National literacy testing programs are an excellent source of essays suitable for calibrating, testing and benchmarking AEG systems. Many countries now have national testing of school students to assess students' learning and abilities in literacy and numeracy. These tests require students to undertake mathematical and literacy testing at age-suitable levels. The literacy tests generally include essay writing. Australia and the United States are two countries which have these tests.

3.3 Literacy and Numeracy Assessment

In Australia and the USA there are three well defined state/national standards for Literacy and Numeracy Assessment, known as Rubrics, namely:

- WALNA - Western Australian Literacy and Numeracy Assessment (Australia),

- NAPLAN - National Assessment Program Literacy and Numeracy (Australia),
- NAEP - National Assessment of Educational Progress (USA).

The Western Australia Department of Education and Training (WADET) assessed primary and secondary students up until 2008 under the Western Australian Literacy and Numeracy Assessment (WALNA) program. These assessments are now conducted Australia wide under the National Assessment Program Literacy and Numeracy (NAPLAN).

3.3.1 WALNA marking criteria

The essays were marked on twelve criteria. Table 3.1, lists these criteria and provides an explanation of each. Each criterion has a number of levels or categories to which the essay can be assigned.

3.3.2 NAPLAN marking criteria

Table 3.2 shows the NAPLAN criteria for assessing student writing.

3.3.3 NAEP marking criteria

The US National Assessment of Educational Progress (NAEP) was a nationwide program under which selected students were tested for writing skills. Eighth grade students were instructed to write on a topic that was classified as either *narrative*, *informative*, or *persuasive* (NAEP, 2010). The detailed Eighth-Grade Informative Writing Scoring Guide is shown in Table 3.3 (NAEP, 2000).

Table3.1. WALNA Marking Criteria (Source: WADET 2006, pp. 7-9)

On Balance Judgement	Holistic judgement of the script.
Spelling	Accuracy of spelling in context of the students' writing, from very familiar and common words to difficult and unusual words. The quality of errors is also taken into account.
Vocabulary	Students' repertoire of words and phrases that they have available for their writing.
Rhetoric Devices	Assessing the quantity, appropriateness AND effectiveness of the device/s used.
Sentence Structure	Sentence completeness, sentence form, variation in beginnings, variety in length, clarity and enhancement of

	meaning, and reading fluency.
Punctuation of Sentences	Capital letters at the start of sentences and full stops, question marks or exclamation marks to finish sentences.
Punctuation within Sentences	Assumes that... students will experiment with internal punctuation before gaining mastery. This experimentation is rewarded in this criterion.
Introduction	Clarity of the writer's position and the degree of direction given to the reader.
Form – Argument and Ideas	Quality of argument; the breadth and quality of ideas and of supporting evidence.
Conclusion	How well the student draws the essay to a close.
Paragraphs	The presence and make-up of paragraphs: whether each point of argument in the essay body is separated by paragraphs; whether topic sentences are used and whether supporting evidence is linked to the topic sentence.
Register	The way speakers and writers adjust the way they speak or write in different social contexts to communicate for different purposes.

Table3.2. NAPLAN Marking Criteria (Source: NAPLAN, 2010b)

Adapted from text in *National Assessment Program – Literacy and Numeracy 2008: Writing – Narrative Marking Guide*. © Australian Curriculum, Assessment and Reporting Authority, 2008.

The material is reproduced with the permission of ACARA.

Audience	The writer's capacity to orient, engage and affect the reader.
Text Structure	The organization of narrative features including orientation, complication and resolution into an appropriate and effective text structure.
Ideas	The creation, selection and crafting of ideas for a narrative.
Character and Setting	Character: The portrayal and development of character. Setting: The development of a sense of a place, time and atmosphere.
Vocabulary	The range and precision of language choices.
Cohesion	The control of multiple threads and relationships over the whole text, achieved through the use of referring words, substitutions, word associations and text connectives.
Paragraphing	The segmenting of text into paragraphs that assists the reader to negotiate the narrative.
Sentence Structure	The production of grammatically correct, structurally sound and meaningful sentences.
Punctuation	The use of correct and appropriate punctuation to aid the reading of the text.
Spelling	The accuracy of spelling and the difficulty of the words used.

Table3.3. NAEP Marking Criteria (Source: NAEP, 2000)

<p>1. Unsatisfactory Response (may be characterized by one or more of the following)</p>	<p>Attempts to respond to prompt, but provides little or no coherent information; may only paraphrase the prompt.</p> <p>Has no apparent organization OR consists of a single statement.</p> <p>Minimal or no control over sentence boundaries and sentence structure; word choice may be inaccurate in much or all of the response.</p> <p>A multiplicity of errors in grammar or usage (such as missing words or incorrect word use or word order), spelling, and punctuation severely impedes understanding across the response.</p>
<p>2. Insufficient Response (may be characterized by one or more of the following)</p>	<p>Presents fragmented information OR may be very repetitive OR may be very undeveloped.</p> <p>Is very disorganized; thoughts are tenuously connected OR the response is too brief to detect organization.</p> <p>Minimal control over sentence boundaries and sentence structure; word choice may often be inaccurate.</p> <p>Errors in grammar or usage (such as missing words or incorrect word use or word order), spelling, and punctuation interfere with understanding in much of the response.</p>
<p>3. Uneven Response (may be characterized</p>	<p>Presents some clear information, but is list-like, undeveloped, or repetitive OR offers no</p>

by one or more of the following)	<p>more than a well-written beginning.</p> <p>Is unevenly organized; the response may be disjointed.</p> <p>Exhibits uneven control over sentence boundaries and sentence structure; may have some inaccurate word choices.</p> <p>Errors in grammar, spelling, and punctuation sometimes interfere with understanding.</p>
4. Sufficient Response	<p>Develops information with some details.</p> <p>Organized with ideas that are generally related, but has few or no transitions.</p> <p>Exhibits control over sentence boundaries and sentence structure, but sentences and word choice may be simple and unvaried.</p> <p>Errors in grammar, spelling, and punctuation do not interfere with understanding.</p>
5. Skilful Response	<p>Develops and shapes information with details in parts of the response.</p> <p>Is clearly organized, but may lack some transitions and/or have occasional lapses in continuity.</p> <p>Exhibits some variety in sentence structure and some good word choices.</p> <p>Errors in grammar, spelling, and punctuation do not interfere with understanding.</p>
6. Excellent Response	<p>Develops and shapes information with well-chosen details across the response.</p>

	<p>Is well organized with strong transitions.</p> <p>Sustains variety in sentence structure and exhibits good word choice.</p> <p>Errors in grammar, spelling, and punctuation are few and do not interfere with understanding.</p>
--	---

3.3.4 The common features of marking criteria (Rubrics)

There are common assessment criteria amongst the three testing programs discussed above. Table 3.4 compares these three assessment rubrics.

Table 3.4. Comparisons of WALNA, NAPLAN, and NAEP Marking Criteria

Criteria	WALNA	NAPLAN	NAEP
On Balance Judgement	Yes		
Spelling	Yes	Yes	Yes
Vocabulary	Yes	Yes	Yes
Rhetoric Devices	Yes		
Sentence Structure	Yes	Yes	Yes
Punctuation of Sentences	Yes	Yes	Yes
Punctuation within Sentences	Yes	Yes	
Introduction	Yes		

Form – Argument and Ideas	Yes	Yes	Yes
Conclusion	Yes		
Paragraphs	Yes	Yes	
Register	Yes		
Audience		Yes	
Text Structure		Yes	
Character and Setting		Yes	
Cohesion		Yes	
Organisation			Yes

The common criteria are related to technical aspects of an essay, with the exception of Argument and Ideas. These technical aspects in general are suited to automated evaluation and scoring, and some of these aspects are incorporated in the system discussed in this exegesis.

3.4 Challenges for Automated Assessment

In chapter 2, four major issues associated with the existing approaches and technologies have been identified, and these are:

1. Many AEG systems tended to derive scores from essay features, and not necessarily from a true understanding of the essay structure and content. We therefore need a new solution to content assessment.
2. Most AEG systems require either training on hundreds of human graded essays on the specific topic, or manual construction of a knowledge repository. When the number of essays to be graded is small, i.e. less than a few hundred, then this training task is not economical to undertake. Because of this, AEG systems are really only suitable when hundreds, or thousands of essays are to be processed. We need an alternate solution to achieve automated grading without the need for a large number of training essays.

3. AEG systems that use multiple linear regression techniques for building a scoring equation assume that there is a linear relationship between the essay features and a score. This is not necessarily true. It may be the case that assuming non-linearity leads to better outcomes, which of course requires different mathematical techniques, such as non-linear regression, or neural networks. This issue is not addressed in this exegesis, but it is being addressed in further research discussed under future work in chapter 8.
4. There were also the four main sources of error in these AEG approaches identified by Valenti et al. (2003) and described in section 2.12 relating to spelling, sentence structures, qualifications and marking templates (Valenti et al., 2003).

These have to be addressed in any cutting edge solution. The system discussed in this exegesis incorporates processes for assessing spelling, content and marking templates.

3.5 Challenges for Automated Literacy Assessment

The marking criteria for human grading of essays in commonly used national testing processes have been explained in section 3.3. There is much in common amongst them. Common to all three rubrics are

1. Spelling,
2. Vocabulary,
3. Sentence Structure,
4. Punctuation, and
5. Argument and/or Ideas.

Items (1) to (4) are easy to automate, however, (5) presents challenges and involves semantic reasoning. This exegesis discusses how these challenges were addressed when the MarkIT AEG system was designed, built and tested. The major features that can be automatically assessed for the quality of an essay are:

1. stylistic aspects,
2. content,
3. structure and organization,
4. mechanistic aspects.

3.5.1 Stylistic aspects

Writing style in an essay can be detected through the use of a word processor. For example, Microsoft Word® can detect a passive sentence in a document and suggest an active sentence to

replace it. The number of passive sentences are recorded by Microsoft Word®, and if these counts exceed a predetermined threshold, a lower mark can be given to the essay if desired.

3.5.2 Content aspects

If the necessary and expected material is covered sufficiently then a high score can be assigned to this feature. Typically, essay term vectors enable this to be easily measured.

3.5.3 Structure and organization

Here the system needs to detect an introduction, and a set of ideas argued for in a satisfactory way. To a certain extent, these items can be measured by looking for key words such as ‘introduction’ or equivalent, and concepts typically associated with structure such as ‘it is argued’, ‘therefore’, ‘in conclusion’ and similar key words or phrases.

3.5.4 Mechanistic aspects

Spelling and grammar are generally very easy to detect in an electronic document, as word processing software includes this functionality as a standard feature.

3.6 Research Questions

This research questions which have guided this long term project are:

1. Could a new solution to content assessment be developed, based on a true understanding of the essay structure and content?
2. Could a software system be developed to assess an essay against specified grading criteria and predict an accurate score for an essay?
3. Could such a system determine the significant grading criteria?
4. Could the system provide feedback to the teacher and student?
5. How would such a system be built?

The answers are complex and will be elaborated in chapters 3 - 8.

3.7 Research Issues

In order for an analysis system to “understand” an essay, and grade it, an intelligent AEG system needs to understand the structure of the essay sentences and phrases.

A review of relevant literature was undertaken to see what others had done in AEG research in order to get an understanding of the field, and the problems and issues others had to deal with. The review, reported in Williams (2001), revealed the early work undertaken by Page in the 1960s which

was discussed in Chapter 2. Details of a number of other systems that were starting to emerge were also discovered and these have been discussed in Chapter 2.

An analysis of the theoretical underpinnings of these systems, and their reported performance, was then undertaken. Details of these systems were discussed in Chapter 2. As discussed in Williams (2001), in general, the correlation between human and computer grades ranged between 0.38 and 0.88. Humans generally agree on grades amongst themselves in the range of 0.75 to 0.90. It appeared that E-rater performed best in regard to human graders, then, in order of best to worst, IEA, TCT and PEG. This research indicated that AEG systems were emerging as useful systems, and could in many cases perform as well as humans in the grading process. From this evaluation it was determined that there was no single successful technological approach to AEG, and that the performances of these systems, when measured against human graders, varied. Technologies that were effective were noted and these influenced the thinking of how the building of an effective system could be undertaken.

3.7.1 Practical limitations of Context Free Phrase Structure Parsers

Many Natural Language Processing (NLP) systems use some kind of a parser to extract the syntax of sentences in a document as an initial step prior to further processing. Semantic analysis then follows. The use of Context Free Phrase Structure Grammar (CFPSG) parsers is commonly suggested in the literature. They require an extensive set of grammar rules which define legitimate syntax structures. However, it is virtually impossible to build a set of grammar rules for free unseen text in practice, because thousands of grammar rules are typically required, and over-generation of possible parse trees results. Increases in parsing time becomes exponential as the parse trees proliferate. So CFPSG parsing cannot be used in all but simple toy documents i.e. simple sentences based on a simple set of grammar rules.

3.7.2 Efficient parsing of sentences

Useful preliminary linguistic computation can be undertaken with less structured parsing. Phrase processing can be an effective alternative to full parsing at the initial processing stage. Phrase processing has the advantage that it does not require an extensive set of grammar rules – a few simple rules suffice. Specifically the process breaks a sentence into syntactically structured components representing noun clauses or phrases and verb clauses or phrases. Often, these structures are sufficient as a preliminary to further processing.

3.7.3 Structures to hold semantic details

For this research, unique structures were developed for noun phrases and verb clauses to hold a semantic representation of an essay. These structures consist of part of speech (POS) descriptors applicable to each type of structure. Words in a sentence are tagged with their POS, and the appropriate thesaurus index numbers are determined. A thesaurus index number is the number assigned to a group of words with a common meaning in a thesaurus. This technique allows for multiple word choices when writing a sentence without substantially changing the meaning of the sentence. Different students usually write essays with dissimilar sentence structures, even if the essay topic is the same. These index numbers are stored in the corresponding POS slot in the structures. Full details of these structures are given in chapter 4.

Table 3.5 shows the part of speech descriptors used in the example that follows.

Table 3.5 Part of Speech Descriptors

Mnemonic	Part of Speech
ADJ	adjective
ADV	adverb
DET	determinant
N	noun
NP	noun phrase
V	verb
VC	verb clause

The exact structures of the NP and VC slots are discussed in chapter 4, but to illustrate the concept and to give a practical example, consider the following. A typical sentence would comprise alternating NPs and VCs as shown in table 3.6.

Table 3.6 Sample Noun Phrase and Verb Clause Representation

A typical first NP slot word and numerical contents would be:	A typical first VC slot word and numerical contents would be:	A typical concluding NP slot word and numerical contents would be:
DET ADJ ADJ N	V ADV ADV	DET N
The small black dog	walked slowly down	the street
100 143 97 678	34 987 67	100 234

The numbers are fictitious thesaurus index numbers for the corresponding words.

3.8 Information Systems Research Methodologies

It is important that the research discussed in this exegesis is based on a sound research methodology. This section discusses a number of possible research methodologies for the work. The one considered most appropriate to gain the best research outcomes is also identified. The appropriate domain of knowledge for this research is the field of Information Systems (IS). The selection of the appropriate research methodology for IS research has been the focus of discussion for many years.

Galliers & Land (1987) presented an early taxonomy of IS research approaches in which they defined two categories: Traditional empirical approaches based on observations, and Modes for newer approaches based on interpretations. Galliers later provided more detail on these approaches (Galliers, 1992). Nunamaker and his colleagues proposed a framework as to how IS research could be undertaken through systems development (Nunamaker, et al., 1991).

The features of these approaches are outlined in the following sections. The purpose for doing this is to locate the research presented in this exegesis in the context of research in the IS domain and to align it with an appropriate methodology. The case for the chosen methodology is argued first. The features of the other approaches are then discussed, and the chosen approach compared and contrasted with these, to further support the choice of the adopted methodology.

3.8.1 The methodology used in this research project

Important tasks undertaken in this research were the analysis, design and development of the MarkIT system. These tasks required careful planning and organisation and were guided by a prototyping methodology, which is explained shortly. Over the last sixty years, methodologies for developing information systems have evolved from ad hoc approaches to highly sophisticated approaches. Underlying most of these approaches is the classic Systems Development Life Cycle (SDLC). It consists of five phases, namely

- Project planning,
- Analysis,
- Design,
- Implementation,
- Support.

(Satzinger, et al., 2002).

There are now many variations of the SDLC, one of which is a prototyping methodology. This methodology involves ascertaining user requirements, building a working model, demonstrating it to the client and receiving feedback, and then modifying and improving the system. A number of iterations of this cycle are undertaken until the evolving system is accepted as the production system (Williams, et al., 1997). Prototyping is a preferred methodology when user requirements are initially unclear and the user wants to explore a number of scenarios to see which one is the most effective in solving the client's problem. It was thus appropriate that MarkIT was developed as a number of increasingly better prototypes, as a number of the algorithms used in the system were based on heuristics derived from experimental theories.

3.8.2 Nunamaker's, Chen's, and Purdin's methodology

Nunamaker and his colleagues proposed a framework as to how IS research could be undertaken through systems development (Nunamaker, et al., 1991). The framework consists of five elements, namely:

- Construct a Conceptual Framework,
- Develop a System Architecture,
- Analyse and Design the System,
- Build the (Prototype) System,
- Observe and Evaluate the System.

Figure 3.1 details this methodology.

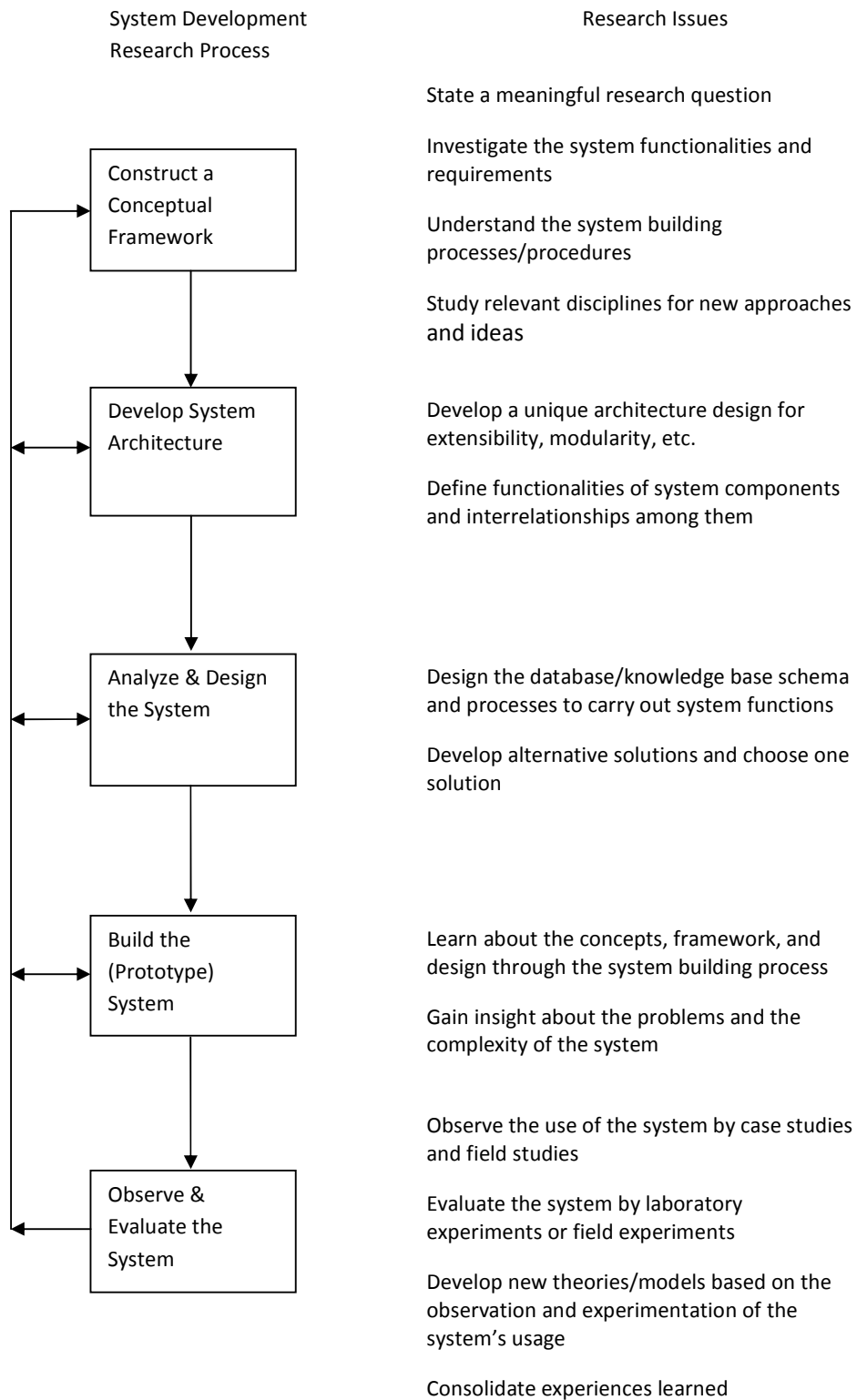


Figure 3.1. A Process for Systems Development Research.

From J.F. Nunamaker, M. Chen, & T.D.M. Purdin, "Systems Development in Information Systems Research," JOURNAL OF MANAGEMENT INFORMATION SYSTEMS, vol. 7, no. 3 (Winter 1991): 89-106, Figure 3. Copyright © 1991 by M.E. Sharpe, Inc. Used by permission.

The authors argued that systems development is a legitimate IS research methodology. It is through the implementation of a system that the underlying theoretical assumptions can be tested for efficacy. They conclude

“Building a system in and of itself does not constitute research. The synthesis and expression of new technologies and new concepts in a tangible product, however, can act as both the fulfilment of the contributing basic research and act as an impetus to continuing research. The important role played by systems development in the life cycle of complex research demonstrates its credibility as a research methodology.”

(Nunamaker et al., 1991, p. 103)

3.8.3 Galliers and Land's traditional empirical approaches to IS research

Galliers & Land (1987) identified seven traditional approaches to IS research. These included

- Theorem proof,
- Laboratory experiment,
- Field experiment,
- Case study,
- Survey,
- Forecasting,
- Simulation.

In the context of IS research, a theorem proof usually involves the use of mathematical techniques to prove the correctness of an algorithm. In the case of this research, the functionality being tested involved many non-algorithmic activities such as constructing the thesaurus database, building a World Wide Web (WWW) presence, and even collecting essays. The theorem proof methodology was thus rejected as being inappropriate to this research.

A laboratory experiment involves testing hypotheses in a controlled environment. The focus is on testing relationships among variables which have been identified in a relevant theory. Mathematical analysis is also generally involved. The research in this instance was much broader in its scope than a simple laboratory experiment. It involved activities such as systems analysis, systems design, programming and testing. It was thus considered that a laboratory experiment would be too narrow in scope to allow for the breadth of the research to be captured.

When a laboratory experiment is taken out of the laboratory and extended into organisations Galliers & Land (1987) call it a field experiment. The field experiment approach was rejected in this research for the same reasons as the laboratory experiment was rejected – the proposed research was much broader in scope than that allowed by a field experiment.

Much of the published research in IS involves case studies. A case study is research conducted in a single or multiple organisation(s) to see whether business activity relationships postulated in theory actually exist in the target organisation(s). Case studies are appropriate when looking for organisational behaviour issues – in this instance the research was concerned with considering essay grading activities from a technical viewpoint and not from an organisational behaviour perspective. This approach was thus considered inappropriate.

A forecast is a prediction about the future behaviour or state of one or many variables of interest. A large amount of data over many time periods is collected and then processed with mathematical statistical methods to arrive at the forecast. Forecasting techniques were used in this research in that the essay grades produced by the AEG system were calculated using multiple linear regression. However this was limited to the testing phase of the research project.

Simulation involves the use of random variables in order to copy the behaviour of a system. It is used when it is too difficult to produce the behaviour of the system without actually building it. Simulation studies undertaken before construction of a system allow engineers to see if the proposed system will perform as expected. Of particular interest are performance measures such as time and cost. This methodology was not considered appropriate here because the effort to build a simulation of the system would be nearly the same as building an operational system.

3.9 Choice of Appropriate Methodology

The MarkIT system was developed and tested as follows. First the underlying theory of representing the meaning of a document was developed (see section 4.4). Secondly, this theory was developed into a set of algorithms. Thirdly these algorithms were coded in a succession of prototype systems, which were tested for accuracy of grading when compared to human graders. Some statistical analysis of the test results was then undertaken. This sequence of events fits very closely to the Nunamaker et al. (1991) methodology explained in section 3.8.2.

3.10 Summary

This sound and reputable research required a solid theoretical foundation that would be accepted in the appropriate discipline. This chapter has surveyed many of the accepted Information Systems

research approaches. A wide variety of methodologies were identified, and the chapter discussed them in detail. The chosen methodology for the development of the MarkIT system fits very closely to the Nunamaker et al. (1991) methodology explained in section 3.8.2. As a result, the adherence to this methodology led to a successful outcome for the research and its artefacts. In chapter 4 a discussion of the concepts of noun phrases and verb clauses is given. An introduction to a modification of document term vectors known as normalised word vectors which are used in the new system MarkIT, is also provided.

Chapter 4 Theoretical Framework of the MarkIT System is unable to be reproduced here because of proprietary reasons.

Chapter 5 The Design and Implementation of MarkIT is unable to be reproduced here because of proprietary reasons.

Chapter 6 Trial, Testing and Evaluation

6.1 Introduction

A prototype system to automatically grade essays based upon the theoretical constructs described in chapter 4 was initially developed. This prototype was developed using the C++ programming language and the Microsoft Windows® platform. The purpose of the prototype was to determine if the theoretical ideas developed during the conceptual phase of the research would in fact lead to a useful system. The initial testing of the prototype was encouraging. The system was then completely rebuilt by a professional programmer using the Java language and the Microsoft Windows® platform.

6.2 Preparations for Automated Grading with MarkIT

The steps involved for a teacher who wishes to use the system follow. A topic is set, and arrangements made to have access to human markers. In order to achieve good results from the system, human markers should be given training, using a scoring rubric, to achieve consistency amongst them. The system is dependent upon the quality of the human grades, as these are used to train the system.

Students are informed of the essay topic, and expected answer length in words, along with the submission date and time. Arrangements are made for the students to submit their essays in electronic form via the World Wide Web or email – typically in Microsoft Word® format. Once the essays have been received, human grading of 200 of the essays is undertaken. Multiple human grading of each essay is desirable for better system performance, but it is not essential. Typically three human graders give better results. Several model answers are then chosen from the human graded essays – typically five of the highest scoring essays are used. The system is then trained on the topic by referencing 100 of the human graded essays – this set is known as the training set. After deriving the scoring equation from the training set, the remaining 100 human graded essays are used to validate the scoring equation. This is done by computer scoring each of the 100 validation essays. The computer scores are compared with the human scores, to see that they are close to the human scores, and that the means and standard deviations are similar. Once this confidence is established, the remaining unscored essays are graded. Feedback is then available to the teacher and students via the World Wide Web.

6.3 A Trial of the Intelligent Essay Assessor

Early in the research, in collaboration with Dr. Heinz Dreher of the School of Information Systems at Curtin University in Perth, Western Australia, a decision was made to conduct a trial of an existing AEG system. The purpose of the trial was to gain experience with usage of an AEG system, and to see if there were any deficiencies in the system that could be improved by developing a new system. A grant of \$10,000 was obtained from the Curtin Business School (CBS) to undertake the trial. Students in the unit Information Systems 100 were invited to submit an essay on a topic comparing mainframe, mini and personal computers. Over 500 essays were received electronically from the students. Three hundred of these were then graded once by three humans, each human grading roughly 100 essays. These human graded essays were then sent via email to the IEA personnel in Colorado, USA, and were used to train the IEA on the topic. Shortly after, the remaining un-graded essays were sent to the USA, and graded by the IEA. One week later the results were received. These indicated that the IEA performed very well. However, the test cost over \$11,000, and there was the inconvenience of the human grading of 300 essays, and then having to send them overseas, and a processing time of about 2 weeks. There was at the time, no useful feedback for teachers and students provided by the IEA. In summary the outcomes of the trial were excessive costs, lengthy turn-around time, physical distance, and a certain amount of inconvenience. These outcomes were taken into consideration when planning the features of the to-be-developed system. Costs had to be kept to a few dollars per essay, quick turn-around for processing was important, distance had to be overcome, possibly by having a World Wide Web presence, and ease of use was important.

6.4 Evaluation of the Performance of MarkIT

Trials of the MarkIT system have been conducted since 2004. In a major trial, 391 essays written by Western Australian year 10 students on the topic of the “School Leaving Age” were analysed by the research team, and various combinations of training and validation essay sets were trialled. The performance of MarkIT closely matched that of the human graders. Table 6.1 shows the correlation coefficients of the performance of the system in relation to the human graders for 289 of the “School Leaving Age” essays (not including the 100 training essays, model answer, and one discarded essay which consisted of only two words). AS, JB and JM are the initials of the three human graders.

Table 6.2 shows the mean and standard deviation for each of the human graders and the computer. AS graded lower on average, and had more variation in grades, than the other two humans. The computer had a lower average than JB and JM, but had less variation than all three humans, indicating good consistency in grading and better than the humans in this regard.

Table 6.1. Correlations for Total Score for 289 “School Leaving Age” Essays

	AS	JB	JM	Human Average
AS	1.00			
JB	0.80	1.00		
JM	0.78	0.81	1.00	
Computer	0.70	0.79	0.78	0.81

Table 6.2 Scoring Statistics for 289 “School Leaving Age” Essays

	AS	JB	JM	Human Average	Computer
Mean	29.40	30.80	30.87	30.36	29.68
Std. Deviation	9.52	7.10	7.84	7.58	6.96

Figures 6.1 to 6.3 illustrate graphically the performance of the human markers, AS, JB and JM on the same essays. Note the variation between them, illustrating that humans generally do not agree on essay grades. The maximum possible mark for an essay was 54. The graphs are organised in ascending order of one of the sets of grades, which leads to the other set of grades to be plotted in a jagged manner. This is because of the score variations amongst the human graders, and between the human grades and the corresponding computer grades.

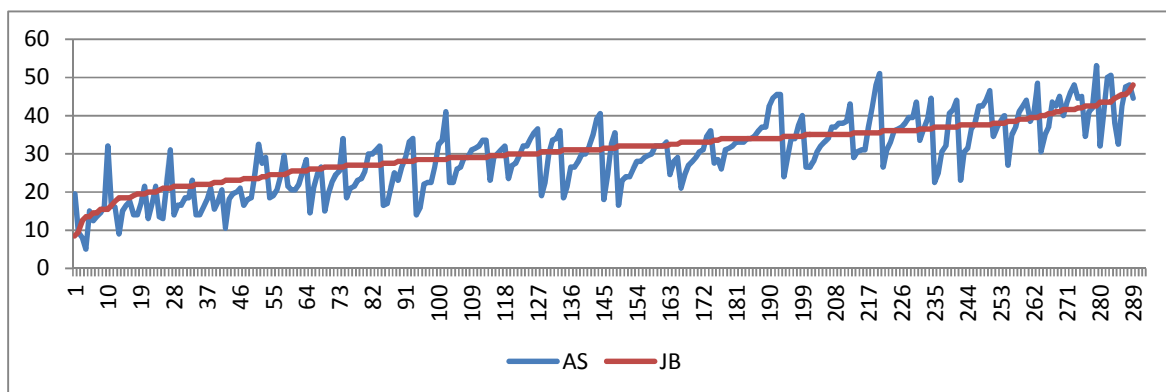


Figure 6.1. Comparison of AS and JB Scores

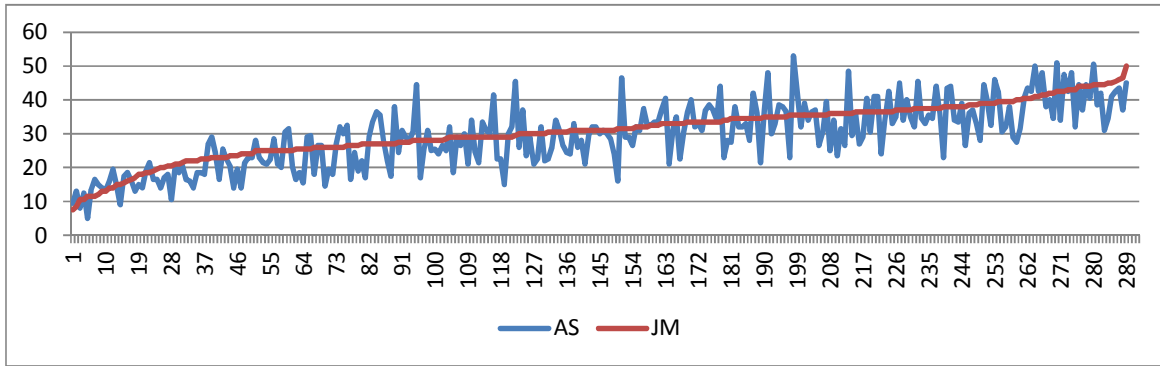


Figure 6.2. Comparison of AS and JM Scores

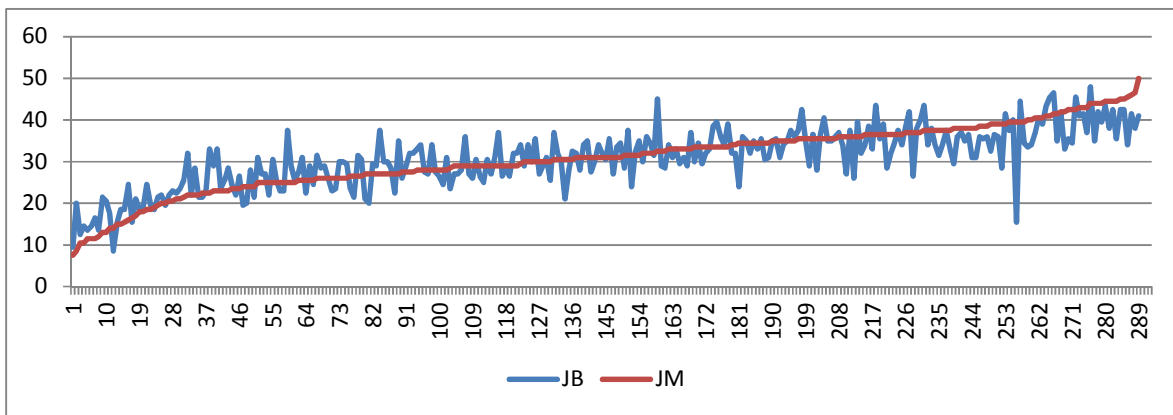


Figure 6.3. Comparison of JB and JM Scores

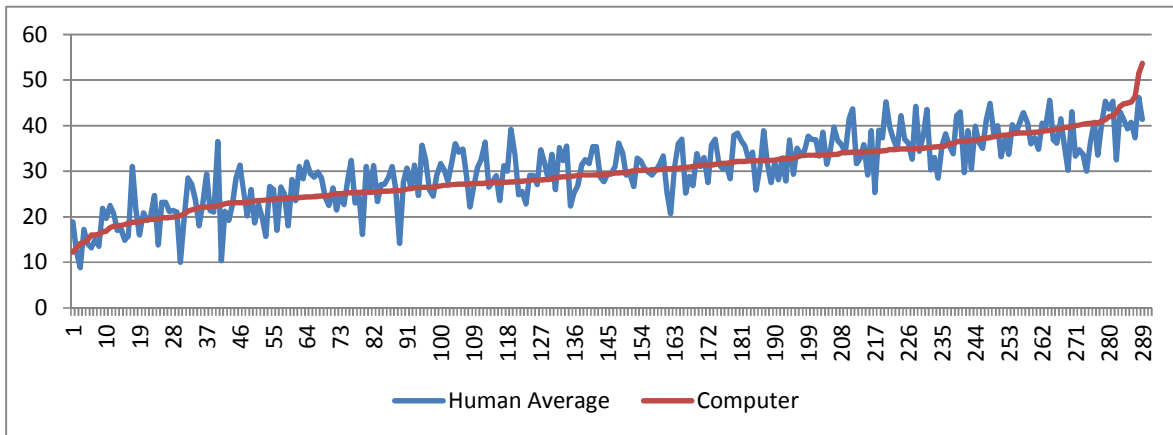


Figure 6.4. Results of Computer Scoring of 289 Essays

Figure 6.4 shows the performance of MarkIT in relationship to the average scores of the three human markers. It can be seen that MarkIT has matched the performance of the humans closely –

the level of variation in the computer assigned scores is the same as that of the humans. The correlation between these scores was 0.81, as shown in Table 6.1.

Figures 6.5 to 6.7 illustrate graphically the performance of the computer against the human markers, AS, JB and JM on the same essays.

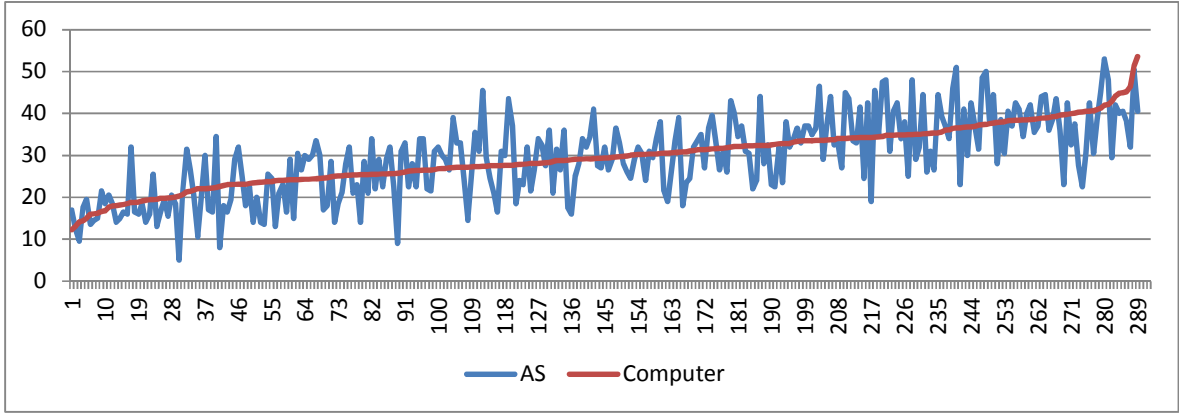


Figure 6.5. Comparison of AS and Computer Scores

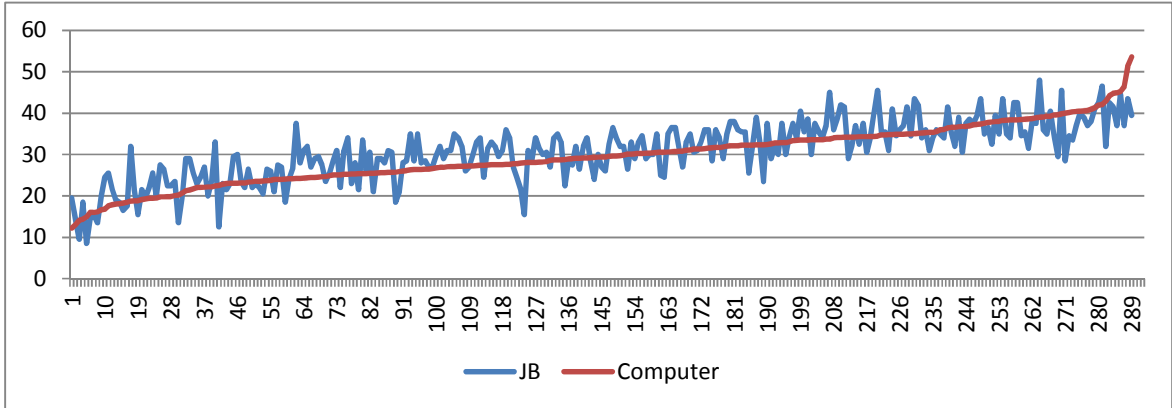


Figure 6.6. Comparison of JB and Computer Scores

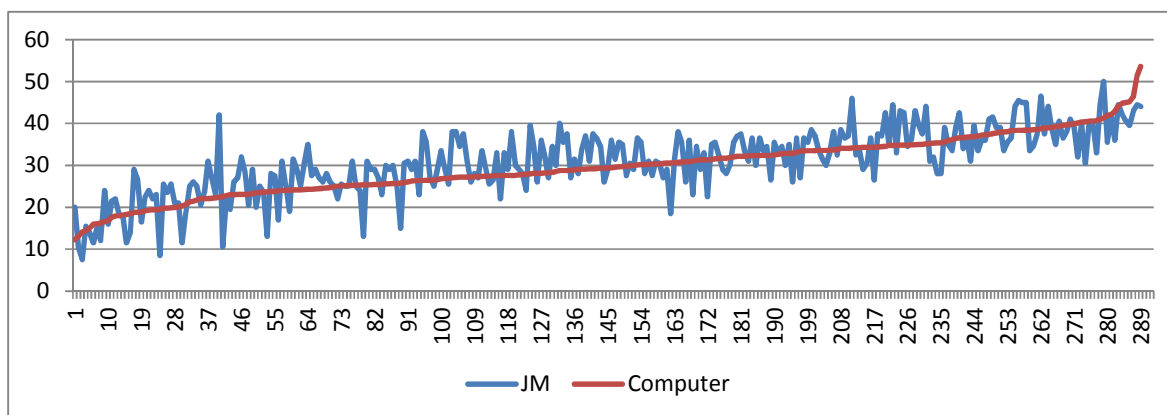


Figure 6.7. Comparison of JM and Computer Scores

Preliminary results of the trial were published in the 2005 report to WADET:

Williams, R. (2005). MarkIT in Action - A Report on the Automatic Grading of DET Year 10 Essays, Curtin University, Perth, Australia.

6.5 Testing Outcomes

The results of testing the new system were provided to WADET, and it was concluded that the new AEG system performed equally as well as the three human markers. The evidence indicates that the theory, and data structures, on which the system is based are sound, and effective in practice.

However, there is a large component of human effort required to set up the grading process. Section 6.6 discusses these issues and other problems encountered during the trials with the WADET essays.

6.6 Challenges Encountered in Testing

Several difficulties were encountered during the testing phase. Firstly, time was spent discussing the essay topic to test. Consideration was given to the number of human graded essays that were available, and the number of essays required for a valid test. Eventually it was decided to use a group of 391 essays on the topic of the school leaving age that had been graded by three humans. These paper-based essays then had to be manually transcribed to Microsoft Word® format, keeping any spelling and grammatical errors made by the students. These essays, together with the human scores, were then provided by WADET for processing. The individual human scores had to be extracted from the spreadsheet files and then averaged. The maximum score that was possible for the chosen scoring criteria was 54. The essay which had the highest human average score was chosen as the model answer. Model answers could also be provided by an examiner if so desired. The first one hundred essays in student identifier order were chosen as the training essays. This was,

in reality, close to a random choice, as there was sufficient variation in the quality of the essays to give a suitable representative set for training. During training, the system can detect an essay that does not have a human score, or a score that does not match a document. This scenario triggers a manual search for the missing document or score, and then restarting the training process when it is resolved.

Sometimes during the linear regression process an error occurs because the data matrix used cannot be inverted. A matrix containing the data derived during the training process is built as part of the process. This matrix must be inverted, a well known mathematical operation, in order to obtain essay feature coefficients for the scoring equation. Technically, not all matrices have an inverse. When this occurs in the MarkIT system, the use of different model answers, or using a different set of training essays, can resolve the problem.

Typically MarkIT can only assign a POS to 85% of the words in an essay – as a result, the system cannot assign a thesaurus concept number to these words. However, this does not impact greatly on the accuracy of the scoring. Multiple POS also lead to issues about how to choose a single POS, and how to choose from multiple thesaurus concept numbers. The choice of a single POS is taken care of during the chunking stage of the algorithm when determining the phrase structures. The choice of a single concept number is resolved by looking at the concepts of the surrounding paragraph, and choosing the one which makes most sense conceptually.

6.7 Summary

MarkIT was tested using 391 essays from year 10 high school students in Western Australia on the topic of the “School Leaving Age”. One hundred human graded essays were selected to train the system. The highest human scored essay was chosen as a model answer. Once the multiple regression analysis was conducted, and a scoring equation developed, it was then applied to the remaining essays to assign a score. The computer scores correlated well with the human scores at .81. The positive outcome led to a major research project with WADET valued at \$740,000 to further develop the system. The next chapter includes an introduction to each of the published papers that form the basis of this exegesis, followed by the papers.

Chapter 7 Publication of Scientific Papers

7.1 Introduction

Six conference papers and one industry report form the basis of this exegesis. The three InSITE conference papers were also published in the online Journal of Issues in Informing Science and Information Technology. Details of these papers are discussed in the following sections.

7.1.1 Automated Essay Grading: An Evaluation of Four Conceptual Models

Williams, R. (2001). Automated Essay Grading: An Evaluation of Four Conceptual Models in Kulski, M. & Herrmann, A. (editors) (2001), *New Horizons in University Teaching and Learning: Responding to Change*, Curtin University, Perth, Australia.

It was important that an understanding of the existing AEG systems be obtained at the start of this research. Four systems were identified in 2001, and this paper discussed the theoretical underpinnings of them, and discusses their performances. A trial of the Intelligent Essay Assessor undertaken by the author and his colleague is discussed. This paper influenced the author in his thinking about possible techniques that would be useful in building a new system. In particular, the ideas relating to normalised word vectors and the use of multiple regression were influential.

7.1.2 Automatically Grading Essays with MarkIT

Williams, R., & Dreher, H. (2004). Automatically Grading Essays with MarkIT. *Journal of Issues in Informing Science and Information Technology*, Vol. 1, pp. 693-700.

The development of the ideas and the corresponding implementation of them took an evolutionary path. This paper was the first paper which discussed some of the ideas of the evolving system, and details of an implementation based on them. The paper discusses a trial of MarkIT with second year law essays. In this case, the scoring equation was handcrafted using some of the features of the essays and one model answer. The final system progressed these ideas into multiple linear regression as a technique for automatically building the scoring equation.

7.1.3 Telecommunications Use in Education to Provide Interactive Visual Feedback on Automatically Graded Essays

Williams, R. & Dreher, H. (2005). Telecommunications Use in Education to Provide Interactive Visual Feedback on Automatically Graded Essays. *Proceedings of International Telecommunications Society Africa-Asia-Australasia Regional Conference*, Perth, Australia.

This paper discussed the MarkIT AEG system with particular emphasis on the interactive visual feedback mechanisms, made possible by its implementation as a World Wide Web application. These mechanisms were made possible by the unique way the system represents essay concepts. The implementation of the system as a web based application was relevant to a telecommunications conference, where the system was presented as a future application where educational services would be delivered over the World Wide Web.

7.1.4 Formative Assessment Visual Feedback in Computer Graded Essays.

Williams, R., & Dreher, H. (2005). Formative Assessment Visual Feedback in Computer Graded Essays. *Journal of Issues in Informing Science and Information Technology*, Vol. 2, pp. 23-32.

The paper discussed the MarkIT system feedback mechanisms in the light of formative assessment. As the research proceeded with the new system, new aspects of the system were explored and documented. The formative feedback consisted of essay mechanical information, such as spelling and grammatical errors, reading ease and reading level, and also comparisons of concept coverage in the student's essay and the model answers. The comparisons were interactive, whereby the teacher and students could explore the alternative ways the essay content could be expressed. The system proved to be useful in providing formative feedback to teachers and students.

7.1.5 MarkIT in Action - A Report on the Automatic Grading of DET Year 10 Essays

Williams, R. (2005). MarkIT in Action - A Report on the Automatic Grading of DET Year 10 Essays, Curtin University, Perth, Australia.

This document reported the findings of a trial of the MarkIT system with WADET and Year 10 student essays on the topic of the School Leaving Age. An important question with the new system was whether it could process a large number of real student essays produced under formal testing conditions. The conclusion of the trial was that they could, and WADET subsequently became an industry partner in a successful Australian Research Council (ARC) Linkage Grant application valued at \$740,000.

7.1.6 The Power of Normalised Word Vectors for Automatically Grading Essays

Williams, R. (2006). The Power of Normalised Word Vectors for Automatically Grading Essays. *Journal of Issues in Informing Science and Information Technology*, Vol. 3, pp. 721-729.

This paper discussed the use of Latent Semantic Analysis (LSA) and Normalised Word Vectors (NWW) for grading essays. It also described further theoretical and practical work that was undertaken to test the author's ideas on an alternative way of grading essays, as opposed to the LSA approach. The

NVW was one of the predictor variables in the regression equation developed for Year 10 essays described in section 7.1.5, and in this case proved to be a useful feature. This was a particularly significant outcome in that the theory of NWV was validated in a practical situation.

7.1.7 A Computational Effective Document Semantic Representation

Williams, R. (2007). A Computational Effective Document Semantic Representation. *Proceedings of IEEE-Digital Ecosystems and Technologies 2007 Conference*, Cairns, Australia, 21-23 February, pp. 410-415.

The paper discussed a technique for representing the semantics of a document. After gaining an understanding of some of the existing AEG systems, and their features and ability to effectively grade real student essays, the author thought that it might be possible to approach the AEG problem in a novel way. This paper described the resultant theory and algorithms. In particular, the development of noun phrase and verb clause structures proved effective in disambiguating multiple thesaurus root concept numbers. Typically, words are classified under a number of root concepts. The system must choose only one of these concepts to store in the structures. This choice is made by considering the surrounding context of the word, in terms of concept numbers. The concept number that is more closely related in meaning to the surrounding concept numbers is chosen. The disambiguated root concept numbers are then used in the NWV.

Chapter 8 Recapitulation and Future Work

8.1 Introduction

This exegesis has described a new and novel way of building an AEG system. The chief contributions to the AEG theory are the normalised word vector, and noun phrase and verb clause structures. The technique for reducing essay words to a thesaurus root concept is also an important technique and has proved effective in practice. This chapter recapitulates the work undertaken for the project, and outlines ideas for further research.

8.2 Overview of the Research

This exegesis discusses the development of a new technique to automatically grade essays. The technique produced improved performance when compared to some existing AEG systems. The development of the new program, known as MarkIT, is also discussed. Its performance for assessing a large number of high school English essays is also described. The system has proved to be successful, and is the subject of a large research collaboration with the Western Australian Department of Education and Training (WADET), where it will be modified to potentially grade thousands of essays each year. The system can grade 400 word essays written in English in under 3 seconds each, and provides comprehensive textual and visual feedback to the teacher and student via the World Wide Web.

8.3 Contribution of the Research

8.3.1 Scientific contribution

The author's approach uses a system of phrase structures developed specifically for the MarkIT system. These structures enable fast chunking of document sentences without the need for extensive grammar rules. Once established, the phrases allow for quick resolution of the context for competing thesaurus concepts for a word, and the corresponding index. It was thus important for this exegesis to examine what other researchers had done in the past in relation to this topic. The approach also makes use of document content vectors known as Normalised Word Vectors. Importantly, trials of the MarkIT system using these technologies and multiple linear regression have shown that the system has a conformance rate with human graders within the variability of the human graders themselves.

8.3.2 Contribution to the education sector and community

The author's approach also addressed the issues discussed in Chapter 3 as follows. Noun phrase and verb clause structures are used to enable fast and effective parsing of sentences, and to have a

simple storage mechanism for sentence semantics. This enables the normalized word vectors to be produced. Model answers are used to address the issues associated with assessing content. Using these features, the MarkIT system was developed to prove the concepts. The system proved successful when tested with real world essays. These contributions were recognized as positive contributions as is evidenced by the awarding of the Australian Research Council Linkage Grant LP0669242 worth a total of \$740,000. The author was one of the Chief Investigators for this grant.

8.3.3 Achieving correlation of 0.81 with human markers

On the trial with WADET Year 10 essays, the system achieved a correlation with the human scores of 0.81, matching the performance of the human graders amongst themselves (Table 6.1). This is surprising in the sense that the topic was the open-ended question on the School Leaving Age in Western Australia, where there was no correct answer.

8.3.4 Integrated summative and formative assessment

The system provides summative assessment in the form of a raw score for an essay, and detailed summative assessment in the form of spelling and grammatical errors, and details of missing content that could be added to an essay to improve its quality. The ability of the system to extract all of this information in seconds saves the teacher from having to spend time doing this manually. The teacher can spend his/her time on higher-level aspects related to the students' learning. This greatly enhances the students' opportunity to learn from the essay writing effort, and improve their performance on essay writing in the future.

8.3.5 Feedback to both the teachers and students

The system is designed so that a teacher and student can sit down at a computer together and explore the outputs of the system for the student's essay. Both can view the system outputs, and this can stimulate discussions about essay features that may be overlooked in a non-computer-based review.

8.3.6 Dynamic and interactive feedback

In the MarkIT system discussed in this exegesis, assessment is not only both summative and formative, but the feedback is dynamic and interactive, which is an innovation in AEG.

Students and teachers can view the student and model answer essays, see comparisons of content, and dynamically view on demand the concepts of interest in both essays. Thesaurus entries can also be viewed for any concept. Concept relationship maps are also provided. These features provide a

rich interactive experience for students so that they can improve their knowledge on the topic, and explore a large number of ways of expressing their knowledge.

8.3.7 Contribution to smart information use as designated by ARC

The Australian Federal Government published Australia's National Research Priorities in 2002 (DEST 2002). The third of four priorities is Frontier Technologies for Building and Transforming Australian Industries. This priority has five goals, the fourth of which is Smart Information Use. The work discussed in this exegesis falls under this goal.

8.3.8 A guide book on technology adoption for the education sector

This exegesis gives a comprehensive review of the current state of the art of the emerging phenomenon of Automated Essay Grading systems. The fact that there are many systems under development, and a number of them are in use in the United States of America, indicates that these systems will play a role in education in the future. It is expected that the systems will be adopted in other countries, particularly in those countries where English is the language of instruction, as most of the published research relates to English language systems. Many AEG systems have been identified, and their features, advantages and limitations are discussed in detail. This exegesis should enable teachers to understand the benefits of using automated grading systems, as well of their pitfalls, and guide them in their choice of a suitable system.

8.3.9 Enhanced teaching and learning experience for both teachers and students

The grading of essays by teachers is very time consuming. Teachers are tempted to limit the number of essay assessments they set for their classes because of the effort involved in grading them. Students then do not get enough practice at the essay writing tasks. AEG systems are now starting to find their ways into the educational sector. They can, in many cases, perform at the same level of accuracy as human graders. However, many have limited feedback to the teachers and students. Feedback is important because students learn from corrections to their grammar, and the provision of alternative ways of expressing ideas. This work is significant in that for the first time an AEG system has been developed that provides interactive visual feedback to the student and teacher, and still performs at the same level as human graders. This will have enormous benefits to teachers and students in that, in a matter of seconds, they can receive a comprehensive analysis of the student essay, and then interact with the feedback to enhance the student's learning experience.

8.3.10 Economic benefit to the Government and education sector

Assessment of large numbers of essays written in English e.g. in hundreds of thousands, is commonplace throughout the world, mainly in system wide standards monitoring. In Australia, it is

estimated that about 1 million essays from years 3, 5, 7, and 9 students are graded each year for standards monitoring. In the United States of America, about 600,000 essays are graded each year for the Graduate Management Admission Test (GMAT). The cost of such programs is significant. The use of automated grading should lead to substantial savings for these projects.

8.3.11 Just in time quality feedback

Many students do not receive feedback in time to learn from it. By the time many university students receive comments on their work, they have moved on to other assessment tasks, or have completed the subject for which they wrote the essay, and do not pay much attention to what is then perceived as of little use. For this reason, timeliness of feedback is very important.

Human generated feedback is quite often based on surface level features of an essay. This is because with high marking loads, teachers quite often cannot spend very much time in assessing individual essays. The MarkIT system provides extensive feedback quickly, and this should enable teachers to focus on the relevance of the feedback, without having to extract the feedback information themselves.

8.4 Limitations of the Research

Since the testing of the system with the WADET essays, other essay topics have been processed by the system. Typically, MarkIT is effective at scoring essays up to 600 words in length. The reason why the performance falls off with essays exceeding 600 words in length is the subject of further enquiry. A possible reason is that the discrimination between essay content obtained by normalising words to their root concept (812 concepts in the Macquarie Thesaurus) falls away as the number of counts for each concept increases. Typically, the system can achieve the level of agreement that human markers achieve amongst themselves. The issue of possible non-linearity of essay features and scores is currently being tackled by the research being conducted under the ARC Linkage Grant discussed in section 8.3.2.

Another limitation of the system is that it has not been tested on creative writing. The nature of the algorithms employed by the system may mean that the system cannot process these types of essays, as it would be hard to provide the necessary model answers. It would also be difficult to find 100 training essays, because of the nature of creative writing, in which the essays would be expected to be different in terms of content and expression.

8.5 Future Directions

The requirement for training essays imposes a limitation on the use of MarkIT in its current form. The system is suitable for large volumes of essays up to 600 words in length. There is considerable interest from textbook publishers for automatically grading short answer questions found at the end of chapters in many books. Publishers obviously do not want to provide 100 human graded essays for every end of chapter question in a book – it would be too cumbersome and expensive to do so. A prototype short answer system has been developed, and initial testing has elicited positive comments from publishers. This system uses some of the technology of MarkIT, but also has new software functionality. It is hoped that this will expand the areas of automated grading.

MarkIT is currently being commercialized, and several organizations have expressed interest in taking it up.

This chapter has outlined the process of setting up a trial of the MarkIT system. Results of testing undertaken with year 10 essays have been presented and discussed. Problems encountered during the testing have been presented and discussed. Overall MarkIT has performed as well as the human graders for essays up to 600 words in length. This indicates that the theoretical foundations of the system are appropriate, and can be incorporated in a piece of software to grade for moderate sized essays.

In Australia, alone the system has the potential to save several million dollars per annum if it were to be adopted by national literacy assessment programs. The ARC grant industry partner is keen to see if a system can be developed that produces scores that follow the Rasch model of Item Response Theory (Bond and Fox, 2007). This theory enables scores to be tested to see if they genuinely assess the abilities that the teachers think they are assessing.

References

- Abney, S P (1991), Parsing by Chunks, in Berwick, R, Abney, S & Tenny, C (eds.), *Principle-Based Parsing*, Kluwer Academic Publishers.
- Baker, F B (1976), Automation of Test Scoring, Reporting, and Analysis, in Thorndike, R L (ed.), *Educational Measurement*, 2nd edn, American Council on Education, Washington, D. C.
- Bloom, B (ed.) (1956), Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook 1, Cognitive Domain, Longmans, Green, New York; Toronto.
- Bond, TG & Fox, CM (2007), *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, Lawrence Erlbaum Associates, Mahwah, N. J.
- Burstein J, Kaplan, R, Wolff, S & Lu, C (1996), Using Lexical Semantic Techniques to Classify Free-Responses, *Proceedings from the SIGLEX96 Workshop*, ACL, University California, Santa Cruz.
- Burstein, J, Kukich, K, Wolff, S, Lu, C & Chodorow, M (1998), Enriching Automated Essay Scoring Using Discourse Marking, Proceedings of the Workshop on Discourse Relations and Discourse Markers, Annual Meeting of the Association of Computational Linguistics, August, Montreal, Canada.
- Burstein, J, Leacock, C & Swartz, R (2001), Automated Evaluation of Essay and Short Answers, in Danson, M (ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough University, Loughborough, UK.
- Cohen, A (2006), Personal email communication with Anna Cohen of WADET, 13 October.
- Christie, JR (1999), Automated Essay Marking – For Both Style and Content, in Danson, M (ed.), *Proceedings of the Third International Computer Assisted Assessment Conference*, Loughborough University, Leicestershire, UK.
- Christie, JR (2003), Automated Essay Marking for Content ~ Does it Work? in Christie, J (ed.), *Proceedings of the Seventh International Computer Assisted Assessment Conference*, Loughborough University, Leicestershire, UK.
- DEST 2002, *Australia's National Research Priorities*, Australian Government Department of Education, Science and Training, viewed 6 October, 2010, <<http://www.dest.gov.au/NR/rdonlyres/61DE0AD2-6857-4D64-9B1D-D42324ACAE55/2850/overview.pdf>>.

Ebel, RL (1979), *Essentials of Educational Measurement*, 3rd edn, Prentice-Hall, Englewood Cliffs, New Jersey.

Flesch, R (1948), A New Readability Yardstick, *Journal of Applied Psychology*, vol.32, pp. 221–233.

Galliers, R (1992), Choosing Information Systems Research Approaches, in Galliers, R (ed.) *Information Systems Research: Issues, Methods and Practical Guidelines*, Blackwell, Oxford, pp. 144-162.

Galliers, R & Land, F (1987), Choosing Appropriate Information Systems Research Methodologies, *Communications of the ACM*, vol. 30, no. 11, pp. 900-902.

Gronlund, NE & Linn, RL (1990), *Measurement and Evaluation in Teaching*, 6th edn, Macmillan, New York.

Hearst, M (2000), The Debate on Automated Essay Grading, *IEEE Intelligent Systems*, vol.15, no. 5, pp. 22-37.

Intelligent Assessment Technologies Ltd (2006), *E-Assessment of Short-Answer Questions*, viewed 6 October, 2010, <<http://www.intelligentassessment.com/>>.

Kolln, M (2004), *Understanding English Grammar*, MacMillan, New York.

Lam H W, Dillon TS & Chang, E (2010), Towards the use of Semi-structured Annotators for Automated Essay Grading, *Proceeding of 24th IEEE International Conference on Advanced Information Network Systems*, IEEE CS Press, Perth, Australia.

Landauer, TK, Foltz, PW & Laham, D (1998), An Introduction to Latent Semantic Analysis, *Discourse Processes*, vol. 25, pp. 259-284.

Larkey, LS (1998), Automatic Essay Grading Using Text Categorization Techniques, *Proceedings of the Twenty First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 90-95.

Lindquist, E F (1955), *Methods and Apparatus for Processing Data*. U.S. Patent 003050248, viewed 6 October, 2010 <<http://www.uspto.gov/patents/process/search/index.jsp>>..

Luhn, HP (1957), A Statistical Approach to Mechanized Encoding and Searching of Literary Information, *IBM Journal of Research and Development*, vol. 1, no. 4 (October).

Macquarie Thesaurus (2010), viewed 8 December 2010,

<http://www.macquarieonline.com.au/thesaurus.html>

Mason, O & Grove-Stephenson, I (2002), Automated Free Text Marking with Paperless School, in Danson, M (ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough University, Loughborough, UK.

Matsumoto, Y & Utsuro, T (2000), Lexical Knowledge Acquisition, in Dale, R, Moisl, H & Somers, H (eds.), *Handbook of Natural Language Processing*, Marcel Dekker, Inc., pp. 563–610.

McGee, T (2006), Taking a Spin on the Intelligent Essay Assessor, in Ericsson, PF & Haswell, RH (eds.), *Machine Scoring of Student Essays: Truth and Consequences*, Logan Utah State University Press, pp. 79-92.

Ming, PY, Mikhailov, AA & Kuan, TL (2000), Intelligent Essay Marking System, in Cheers, C (ed.), *Learners Together*, February, NgeeANN Polytechnic, Singapore.

Mitchell, T, Russell, T, Broomhead, P & Aldridge, N (2002), Towards Robust Computerised Marking of Free-Text Responses, in Danson, M (ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough University, Leicestershire, UK.

NAEP (2000), viewed 6 October, 2010, <<http://nces.ed.gov/pubs2000/2000506.pdf>>.

NAEP (2010), viewed 6 October, 2010,

<<http://nces.ed.gov/nationsreportcard/writing/whatmeasure.asp>>.

NAPLAN (2010a), viewed 6 October, 2010, <<http://www.naplan.edu.au/>>.

NAPLAN (2010b). viewed 6 October, 2010,

<<http://www.naplan.edu.au/verve/resources/napmarkguide08.pdf>>

Nunamaker, JF, Chen, M & Purdin, TDM (1991), Systems Development in Information Systems Research. *Journal of Management Information Systems*, vol. 7, no. 3, pp. 89-106.

Page, EB (1966), The Imminence of Grading Essays by Computer, *Phi Delta Kappan*, January, pp. 238-243.

Page, EB (1994), Computer Grading of Student Prose, Using Modern Concepts and Software, *Journal of Experimental Education*, vol. 62, pp. 127-142.

Page, EB & Petersen, NS (1995), The Computer Moves into Essay Grading, *Phi Delta Kappan*, March, pp. 561-565.

Powers, DE, Burstein, JC, Chodorow, M, Fowles, ME & Kukich, K (2001), *Stumping E-Rater: Challenging the Validity of Automated Essay Scoring*, ETS Research Report 01-03, Educational Testing Service, Princeton, NJ.

Race, P (2005), *Making Learning Happen: A Guide for Post-Compulsory Education*, Sage Publications.

Readability Formulas.Com (2010a), The Flesch Reading Ease Readability Formula, viewed 14 October, 2010,

<<http://www.readabilityformulas.com/flesch-reading-ease-readability-formula.php>>

Readability Formulas.Com (2010b), The Flesch Grade Level Readability Formula, viewed 14 October, 2010,

<<http://www.readabilityformulas.com/flesch-grade-level-readability-formula.php>>.

Rudner, LM & Liang, T (2002), Automated Essay Scoring Using Bayes' Theorem, *Journal of Technology, Learning, and Assessment*, vol. 1, no. 2, viewed 6 October, 2010, <<http://www.itla.org>>.

SAGrader (2010), viewed 6 October, 2010, <<https://www.sagrader.com/login>>.

Salton, G (1966), Automatic Phrase Matching, in Hays, DG (ed.), *Automatic Language Processing*, American Elsevier, New York.

Salton, G (1967), *Information Storage and Retrieval*, Scientific Report No. ISR-12, Department of Computer Science, Cornell University, New York.

Salton, G (1968), *Automatic Information Organization and Retrieval*, McGraw-Hill, New York.

Salton, G (1975a), *A Theory of Indexing*, Society for Industrial and Applied Mathematics, Philadelphia.

Salton, G (1975b), *Dynamic Information and Library Processing*, Prentice-Hall, Englewood Cliffs, New Jersey.

Salton, G & McGill, M (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.

Salton, G (1989), *Automatic Text Processing. The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, Massachusetts.

Satzinger, J, Jackson, R & Burd, S (2002), *Systems Analysis and Design in a Changing World*, 2nd edn, Course Technology Thomson Learning.

Shermis, M & Burstein, J (2003), *Automated Essay Scoring: A Cross-Disciplinary Perspective*, Lawrence Erlbaum Associates, New Jersey, USA.

Siemens (2000), *Method and Arrangement to Determine a Structure of a Plurality of Electronic Data Sets with Regard to an Electronic Thesaurus*, Patent No. DE 198 43 450 A1, German Patent and Trademark Office. (Translated into English on behalf of Griffith Hack Patent Attorneys, Perth, Australia).

Ullman, JD (2007), *Making Education Less Expensive*, Power Point® Presentation provided to the author, 31 May.

Valenti, S, Neri F & Cucchiarelli, A (2003), An Overview of Current Research on Automated Essay Grading, *Journal of Information Technology Education*, vol. 2, pp. 319–330.

Vantage Learning (2004), *Intellimetric Web-based Essay Scoring Engine*, viewed 6 October, 2010, <<http://www.vantage.com/pdfs/intellimetric.pdf>>.

Vantage Learning (2005), How Intellimetric Works, viewed 6 October, 2010, <http://www.vantagelearning.com/docs/intellimetric/IM_How_IntelliMetric_Works.pdf>.

WADET (2006), Writing Marking Guide MSE 2006, Western Australian Department of Education and Training, Perth.

Williams, R (2001), Automated Essay Grading: An Evaluation of Four Conceptual Models, in Kulski, M & Herrmann, A (eds.), *New Horizons in University Teaching and Learning*, Curtin University, Perth, Australia, pp. 173-184.

Williams, R, Klass, D & Morien, R (1997), Developing the ALLOCATE Resource Allocation System: A Case Study in Prototyping a Decision Support System, *Proceedings of the 8th Australasian Conference on Information Systems*, University of South Australia, Adelaide, South Australia.

Glossary of Terms

AEG	Automated Essay Grading
AES	Automated Essay Scoring
ARC Linkage Grant	Australian Research Council grant
AutoMark	An essay grading system
BETSY	Bayesian Essay Test Scoring System
C#	An Object Oriented programming language
C++	An Object Oriented programming language
Chunking	Breaking sentences into surface level phrases
Correlation	A statistical technique to measure the linearity between data points. It ranges between -1 and +1
Cosine	A measure of an angle in a triangle, calculated by the ratio of the length of the adjacent side formed by the length of the hypotenuse
C-Rater	Conceptual Rater. An essay grading system
E-Rater	Electronic Essay Rater. An essay grading system
ETS 1	Educational Testing Service 1. An essay grading system
IEA	Intelligent Essay Assessor. An essay grading system
IEMS	Intelligent Essay Marking System. An essay grading system
Intellimetric	An essay grading system
IS	Information System
IT	Information Technology
Java	An Object Oriented programming language

LSA	Latent Semantic Analysis
MarkIT	An Automated Essay Grading System described in this exegesis
MSE9	Monitoring Standards in Education Year 9
NAEP	US National Assessment of Educational Progress
NAPLAN	National Assessment Program Literacy and Numeracy
NLP	Natural Language Processing - the use of computers to process human languages
NP	Noun Phrase
NWV	Normalised Word Vector. A vector in n-dimensions formed from the counts of document words in each of n categories in a thesaurus
Objective test	A test using multiple choice or true/false questions
Paperless School Free Text Marking Engine	An essay grading system
PEG	Project Essay Grade. An essay grading system.
Phrase Structure	Standard structures used to define noun phrases and verb clauses
POS	Part of Speech
Prototype	In Information systems terms, a system that models the required functionality in order for end users to check the look and feel of a new system
Regression	A mathematical technique to fit a straight line to multiple data points
SAGrader	An essay grading system

SDLC	Systems Development Life Cycle
SEAR	Schema Extract Analyse and Report
	An essay grading system
TCT	Text Categorisation Technique. An essay grading system
VC	Verb Clause
Vector	A mathematical construct, generally represented by a line, for an object that has both direction and magnitude
VP	Verb Phrase
WADET	Western Australian Department of Education and Training
WALNA	Western Australian Literacy and Numeracy Assessment
WWW	The World Wide Web

Appendix A Statements of Co-Authorship

To Whom It May Concern

I, Robert Francis Williams, contributed the content specified below to the following publications:

[1] Williams, R., and Dreher, H. (2004). Automatically Grading Essays with MarkIT. *Journal of Issues in Informing Science and Information Technology*, Vol. 1, pp. 693-700.

I was the lead author of this refereed paper. I wrote the section which outlines the features of MarkIT. I also co-authored the section on the performance of MarkIT. The paper discussed an early version of the MarkIT AEG system, and some results obtained with testing it with a small number of university law essays. Whilst there were discussions and collaboration between the authors, the design of the features described in this paper was undertaken by myself.

[2] Williams, R. and Dreher, H. (2005). Telecommunications Use in Education to Provide Interactive Visual Feedback on Automatically Graded Essays. *Proceedings of International Telecommunications Society Africa-Asia-Australasia Regional Conference*, Perth, Australia.

I was the lead author of this refereed paper. I wrote the sections on the theory of semantic representations and the section on the comparison of MarkIT and the human scores. This paper discussed the MarkIT AEG system with particular emphasis on the interactive visual feedback mechanisms. These mechanisms were made possible by the unique way the system represents essay concepts. Whilst there were discussions and collaboration between the authors, the development of the theoretical ideas described in this paper, apart from some of the visual feedback, was undertaken by myself.

[3] Williams, R., and Dreher, H. (2005). Formative Assessment Visual Feedback in Computer Graded Essays. *Journal of Issues in Informing Science and Information Technology*, Vol. 2, pp. 23-32.

I was the lead author of this refereed paper. I researched and wrote the section of the paper which discusses the production AEG systems. I also wrote the section detailing the features of MarkIT. The paper discussed the MarkIT system feedback mechanisms in the light of formative assessment. As the research proceeded with the new system, new aspects of the system were explored and documented. The system proved to be useful in providing formative feedback to teachers and students. Whilst there were discussions and collaboration between the authors, the development of the theoretical ideas relating to the scoring algorithm described in this paper was undertaken by myself.

R Williams 12/7/10

I, as a Co-Author, endorse that these levels of contributions by the candidate indicated above are appropriate.

H Dreher 12/7/10
Heinz Dreher

Appendix B Copyright Permissions

I warrant that I have obtained, where necessary, permission from the copyright owners to use any third-party copyright material reproduced in this exegesis, or to use any of my own published work in which copyright is held by another party.

Bruce Ridley

From: Bruce Ridley
Sent: Wednesday, 7 April 2010 11:05 AM
To: 'Copyrights@ieee.org'
Subject: RE: Including IEEE papers in Higher Degree Theses

Dear Jacqueline Hansson

Thank you for your favourable response - it is most helpful to have this clarification of the IEEE's position. We don't foresee any difficulty in meeting the IEEE's requirements and will make sure that the appropriate notice or message is inserted whenever IEEE papers are included as part of an HDR thesis.

Kind regards
Bruce Ridley

-----Original Message-----

From: j.hansson@ieee.org [mailto:j.hansson@ieee.org]
Sent: Thursday, 1 April 2010 9:09 PM
To: Bruce Ridley
Subject: Re: Including IEEE papers in Higher Degree Theses

Comments/Response to Case ID: 00247691

ReplyTo: Copyrights@ieee.org

From: Jacqueline Hansson

Date: 04/01/2010

Subject: Re: Including IEEE
papers in Higher
Degree Theses

Send To: "Bruce Ridley"
<B.Ridley@exchange.curtin.edu.au>

cc:

Dear Bruce Ridley :

Yes, your students may reprint their IEEE copyrighted papers in their theses and, if you wish, have their theses include IEEE copyrighted material placed on their university's website. We have only two requirements that must be satisfied before they can do so.

(1) The following copyright/credit notice must appear prominently either on the first page of the reprinted material or prominently in the references of the reprinted paper, with the appropriate details filled in:
© [year] IEEE. Reprinted, with permission, from [IEEE publication title, paper title, and author names].

(2) Additionally, if the thesis is to appear on the university's website, the following message should be displayed either at the beginning of the credits or in an appropriate and prominent place on the website: This material is posted here with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of the Curtin University of Technology's products or services. Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by writing to pubs-permissions@ieee.org. By choosing to view this material, you agree to all provisions of the copyright laws protecting it.

Sincerely,

Jacqueline Hansson

Bob Williams

From: Eli Cohen [EliCohen@InformingScience.org]
Sent: Friday, 19 October 2007 10:28 AM
To: Robert Williams
Subject: RE: Permission to reproduce 4 InSITE articles

Bob,

After you get this project done, let me know if you want ISI to publish it as a book. As you know we publish academic works related to informing science that have no or little commercial market. We place the book online for free access to everyone and also publish it in print so you can buy a copy for your parents' bookshelf.

All the best,

--eli

-----Original Message-----

From: Robert Williams [mailto:Bob.Williams@cbs.curtin.edu.au]
Sent: Thursday, October 18, 2007 6:59 PM
To: EliCohen@InformingScience.org
Subject: RE: Permission to reproduce 4 InSITE articles

Thank you Eli.

Your prompt reply is appreciated.

Bob

Regards

Robert F Williams
Lecturer
School of Information Systems
Curtin University of Technology
GPO Box U1987
Perth
Western Australia 6149
Australia
Tel +61 8 9266 7102
Fax +61 8 9266 3076
Email bob.williams@cbs.curtin.edu.au
www.essaygrading.com

-----Original Message-----

From: Eli Cohen [mailto:EliCohen@InformingScience.org]
Sent: Friday, 19 October 2007 12:24 AM
To: Robert Williams
Subject: RE: Permission to reproduce 4 InSITE articles

Bob,

As the author, ISI provides you with full re-use of your material published with us, even for profit. (We just advise people if they re-use their prior work without citing their prior publication, someone may accuse them of self-plagiarism. This does not appear to be your intent.)

In any case, we provide everyone free reprint of the material for classroom use, as is shown on the copyright statement on each article.

But it is always good to have an email giving permission in your CYA (Cover Your Behind) file.

All the best,

-eli

-----Original Message-----

From: Robert Williams [mailto:Bob.Williams@cbs.curtin.edu.au]

Sent: Thursday, October 18, 2007 1:39 AM

To: Eli Cohen

Subject: Permission to reproduce 4 InSITE articles

Hi Eli

The following papers have been presented at InSITE conferences by myself or a colleague:

Palmer, J., Williams, R., and Dreher, H. (2002). Automated Essay Grading System Applied to a First Year University Subject - How can we do it better? Proceedings of Informing Science and IT Education Joint Conference, Cork, Ireland, pp. 1221-1229.

Williams, R., and Dreher, H. (2004). Automatically Grading Essays with MarkIT. Journal of Issues in Informing Science and Information Technology, Vol. 1, pp. 693-700.

Williams, R., and Dreher, H. (2005). Formative Assessment Visual Feedback in Computer Graded Essays. Journal of Issues in Informing Science and Information Technology, Vol. 2, pp 23-32.

Williams, R. (2006). The Power of Normalised Word Vectors for Automatically Grading Essays. Journal of Issues in Informing Science and Information Technology, Vol. 3, pp 721-729.

I am currently writing up my essay grading work as a PhD exegesis - basically tying together 8 articles into a cohesive thesis.

I would like to reproduce the above articles in the exegesis.

Can you please grant me permission to do so? The exegesis is not for commercial purposes.

Thank you

Regards

Robert F Williams
Lecturer
School of Information Systems
Curtin University of Technology
GPO Box U1987
Perth
Western Australia 6149
Australia
Tel +61 8 9266 7102
Fax +61 8 9266 3076
Email bob.williams@cbs.curtin.edu.au
www.essaygrading.com

Bob Williams

From: Gary Madden
Sent: Thursday, 18 October 2007 4:50 PM
To: Robert Williams
Subject: RE: Permission to reproduce article

Bob -

No problem permission granted, let me know if I can help in any other way.

Best,

Gary

-----Original Message-----

From: Robert Williams
Sent: Thursday, 18 October 2007 4:47 PM
To: Gary Madden
Subject: Permission to reproduce article

Hi Gary

I presented the following paper at the ITS conference in 2005, and it is published in the conference proceedings CD:

Williams, R. and Dreher, H. (2005). Telecommunications Use in Education to Provide Interactive Visual Feedback on Automatically Graded Essays. Proceedings of International Telecommunications Society Africa-Asia-Australasia Regional Conference, Perth, Australia.

I am writing up my essay grading research as a PhD by Supplication - basically tying all the threads from 8 papers together in an exegesis.

Can you please grant me permission to reproduce the above article in the exegesis?

Thank you

Bob

Regards

Robert F Williams
Lecturer
School of Information Systems
Curtin University of Technology
GPO Box U1987
Perth
Western Australia 6149
Australia
Tel +61 8 9266 7102
Fax +61 8 9266 3076
Email bob.williams@cbs.curtin.edu.au
www.essaygrading.com

Bob Williams

From: EDcontactcenter [EDcontactcenter@edpubs.gov]
Sent: Monday, 19 July 2010 9:36 PM
To: Bob Williams
Subject: RE: Permission to reproduce table

Dear Customer,

Thank you for interest in U.S. Department of Education Publication Center, unless specifically stated otherwise, all **publications** issued by the U.S. Department of Education (ED) and all information available on ED's website www.ed.gov and <http://edpubs.ed.gov> are in the **public** domain. These **publications** and information may be **reproduced** for non-commercial purposes without prior consent (with attribution to the U.S. Department of Education or the appropriate source).

Sincerely,

Customer Service

US Department of Education Publication Center

1-877-433-7827

<http://edpubs.ed.gov>

Always remember to include original message text with your reply

From: Bob Williams [mailto:Bob.Williams@cbs.curtin.edu.au]
Sent: Monday, July 19, 2010 4:32 AM
To: edpubs@inet.ed.gov
Subject: Permission to reproduce table

Hi

I am writing up research on automated essay grading for a PhD.

I would like permission to reproduce the table

"Eight-Grade Informative Writing Scoring Guide"

Published online in NAEP FACTS Vol 5 No 2 November 2000 page 2

Can you please grant this permission?

Thank you

Best regards

Robert F Williams
Lecturer
School of Information Systems
Curtin University of Technology
GPO Box U1987
Perth

Bob Williams

From: Elizabeth Granda [egranda@mesharpe.com]
Sent: Thursday, 18 October 2007 10:44 PM
To: Robert Williams
Subject: Permission to use figure

18 October 2007

TO: Robert Williams
FR: Elizabeth Granda, Rights & Permissions, M.E. Sharpe, Inc.
RE: Permisson to use material from JMIS

We are pleased to grant permission to you for the reprinting of: J.F. Nunamaker, M. Chen, & T.D.M. Purdin, "Systems Development in Information Systems Research," JOURNAL OF MANAGEMENT INFORMATION SYSTEMS, vol. 7, no. 3 (Winter 1991): 89-106, Figure 3 - "A Process for Systems Development Research" (hereafter referred to as "the Material")

For use in: Robert Williams, PhD dissertation (Fall-Winter 2007)

This permission is subject to the following conditions:

1. Payment on or before initial publication of the Material of the following fee: US\$ NO FEE.
2. Each copy containing our Material that is reproduced or distributed MUST BEAR the following copyright notice:

From J.F. Nunamaker, M. Chen, & T.D.M. Purdin, "Systems Development in Information Systems Research," JOURNAL OF MANAGEMENT INFORMATION SYSTEMS, vol. 7, no. 3 (Winter 1991): 89-106, Figure 3. Copyright © 1991 by M.E. Sharpe, Inc. Used by permission.

3. Permission is granted for non-exclusive world rights in the English language only.
4. Permission is granted for print usage only. No permission is granted for any other uses, including but not limited to granting others permission to photocopy or otherwise reproduce the Material via print, electronic, or other media.
5. This permission grant is for publication of one edition of the book (including non-profit editions for the visually handicapped), with a print quantity not exceeding 5000 copies.
6. This permission does not apply to any material copyrighted by or credited in our publication to another source. It is the responsibility of the prospective publisher for determining the source of any such material.

With best wishes,
Elizabeth Granda
~ ~ ~ ~ ~
Elizabeth Granda
Associate Editor and
Rights & Permissions Manager
M.E. SHARPE, INC.
80 Business Park Drive
Armonk, NY 10504 USA

Bob Williams

From: Tina Kulski
Sent: Friday, 19 October 2007 9:01 AM
To: Robert Williams
Cc: 'Allan Herrmann'
Subject: RE: Permission to reproduce article from TLF 2001

Hi Robert

Permission granted - I've ccd the co-editor (Allan Herrmann) in the email for his information.

All the best with your thesis

Cheers
Tina

Associate Professor Tina Kulski
Office of Teaching & Learning
Curtin University of Technology
GPO Box U1987, Perth, WA 6845
CRICOS Provider Code 00301J

The information contained in this email and any attached files is strictly private and confidential. This email should be read by the intended addressee only. If the recipient of this message is not the intended addressee, please re-send it to the sender at Curtin University of Technology and promptly delete this email and any attachments. The intended recipient of this email may only use, reproduce, disclose or distribute the information contained in this email and any attached files with Curtin's permission. If you are not the intended addressee, you are strictly prohibited from using, reproducing, disclosing or distributing the information contained in this email and any attached files. Curtin advises that this email and any attached files should be scanned to detect viruses. Curtin does not represent that this email including any attachments is free from computer viruses or other faults or defects. Curtin will not be liable to you or to any other person for any loss or damage (including direct, consequential or economic loss or damage) however caused and whether by negligence or otherwise which may result directly or indirectly from the use of this email or any files attached to this email. It is the responsibility of the person opening any files attached to this email to scan those files for computer viruses

-----Original Message-----

From: Robert Williams
Sent: Thursday, 18 October 2007 4:30 PM
To: Tina Kulski
Subject: Permission to reproduce article from TLF 2001

Hi Dr Kulski

I had the following article published by you as an editor:

Williams, R. (2001). Automated Essay Grading: An Evaluation of Four Conceptual Models. In Kulski, M. and Herrmann, A. (editors) *New Horizons in University Teaching and Learning*. Curtin University of Technology, Perth, Australia, pp. 173-184.

I am writing up my essay grading research as a PhD by Supplication - basically tying all the threads from 8 papers together in an exegesis.

Can you please grant me permission to reproduce the above article in the exegesis?

Thank you

Regards

Robert F Williams
Lecturer

Bob Williams

From: Jeffrey Ullman [ullman@gmail.com]
Sent: Friday, 19 October 2007 12:22 AM
To: Robert Williams
Subject: Re:

Sure; feel free to use them. Just put my name on them.

regards.
---jdu

On 10/18/07, Robert Williams <Bob.Williams@cbs.curtin.edu.au> wrote:

> Hi Professor Ullman
>
> Some time ago you sent me the attached Power Point slides.
>
> I am currently writing up some research in automated essay grading as
> a PhD exegesis.
>
> I would like to reproduce the tables in slides 3 and 5 in my exegesis.
>
> Can you please give me permission?
>
> Thank you.
>
> Regards
>
> Robert F Williams
> Lecturer
> School of Information Systems
> Curtin University of Technology
> GPO Box U1987
> Perth
> Western Australia 6149
> Australia
> Tel +61 8 9266 7102
> Fax +61 8 9266 3076
> Email bob.williams@cbs.curtin.edu.au
> www.essaygrading.com
>
>
>

Bob Williams

From: Marilyn McKee [Marilyn.McKEE@det.wa.edu.au]
Sent: Tuesday, 13 July 2010 2:08 PM
To: Bob Williams
Subject: RE: Permission to use essay grading report

Hi Bob

I am happy to give you permission to use the report entitled "MarkIT in Action" in your PhD exegensis.

Kind regards

Marilyn

Marilyn McKee
A/Manager, Educational Measurement Branch
Department of Education

Phone: (08) 9264 4508 Fax: (08) 9264 4045
Email: Marilyn.McKEE@det.wa.edu.au

From: Bob Williams [mailto:Bob.Williams@cbs.curtin.edu.au]
Sent: Monday, 12 July 2010 12:50 PM
To: marilyn.mckee@det.wa.edu.au
Subject: Permission to use essay grading report

Hi Marilyn

Thank you for taking my call today.

I produced the attached report entitled "MarkIT in Action" in 2005.

It details the results of using the MarkIT automated essay grading system with Year 10 student essays on the topic of "The School Leaving Age".

I wish to include this report in my PhD exegesis and seek your permission to do so.

If you are willing to grant me permission, can you please do so by replying to this email?

Thank you

Bob

Best regards

Robert F Williams
Lecturer
School of Information Systems
Curtin University of Technology
GPO Box U1987
Perth
Western Australia 6845
Australia
Tel +61 8 9266 7102
Fax +61 8 9266 3076
Email bob.williams@cbs.curtin.edu.au
www.essaygrading.com

Bob Williams

From: Marilyn McKee [Marilyn.McKee@det.wa.edu.au]
Sent: Tuesday, 20 July 2010 3:48 PM
To: Bob Williams
Subject: RE: Permission to use table

Hi Bob

You have permission to use the table below from the Writing Marking Guide MSE 2006. Western Australian Department of Education and Training, Perth, pages 7-9 in your PhD publication.

Kind regards

Marilyn

Marilyn McKee
A/Manager, Educational Measurement Branch
Department of Education

Phone: (08) 9264 4508 Fax: (08) 9264 4045
Email: Marilyn.McKee@det.wa.edu.au

From: Bob Williams [mailto:Bob.Williams@cbs.curtin.edu.au]
Sent: Tuesday, 20 July 2010 3:36 PM
To: Marilyn McKee
Subject: Permission to use table

Hi Marilyn

Further to my request last week regarding my PhD publication, is it possible to obtain permission to reproduce the following table adapted from

WADET (2006). Writing Marking Guide MSE 2006. Western Australian Department of Education and Training, Perth, pages 7-9?

On Balance Judgement	Holistic judgement of the script
Spelling	Accuracy of spelling in context of the students' writing, from very familiar and common words to difficult and unusual words. The quality of errors is also taken into account
Vocabulary	Students' repertoire of words and phrases that they have available for their writing
Rhetoric Devices	Assessing the quantity, appropriateness AND effectiveness of the device/s used
Sentence Structure	Sentence completeness, sentence form, variation in beginnings, variety in length, clarity and enhancement of meaning, and reading fluency

Punctuation of Sentences	Capital letters at the start of sentences and full stops, question marks or exclamation marks to finish sentences
Punctuation within Sentences	Assumes that... students will experiment with internal punctuation before gaining mastery. This experimentation is rewarded in this criterion
Introduction	Clarity of the writer's position and the degree of direction given to the reader
Form – Argument and Ideas	Quality of argument; the breadth and quality of ideas and of supporting evidence
Conclusion	How well the student draws the essay to a close
Paragraphs	The presence and make-up of paragraphs: whether each point of argument in the essay body is separated by paragraphs; whether topic sentences are used and whether supporting evidence is linked to the topic sentence
Register	The way speakers and writers adjust the way they speak or write in different social contexts to communicate for different purposes

Thank you

Bob

Best regards

Robert F Williams
Lecturer
School of Information Systems
Curtin University of Technology
GPO Box U1987
Perth
Western Australia 6845
Australia
Tel +61 8 9266 7102
Fax +61 8 9266 3076
Email bob.williams@cbs.curtin.edu.au
www.essaygrading.com

Our ref: AIC01513

3 September 2010

Robert Williams
School of Information Systems
Curtin University of Technology
GPO Box U1987
Perth
WA 6845
Email: bob.williams@cbs.curtin.edu.au

Dear Bob

Permission to use specified Material

1. We refer to your request of 19 July 2010 to use the material described or attached in the Annexure to this letter (the **Licensed Material**) in connection with the purposes listed in the Annexure.

Definitions

2. For the purposes of this letter:
 - **ACARA** means the Australian Curriculum, Assessment and Reporting Authority, a body corporate established by section 5 of the *Australian Curriculum, Assessment and Reporting Authority Act 2008* (Cth) ABN 54 735 928 084;
 - **Licence** means the rights granted to the Licensee in paragraph 3; and
 - **Licensee** means the addressee listed above.

Licence

3. ACARA grants to the Licensee a non-transferable, non-exclusive licence to:
 - a. use the Licensed Material;
 - b. reproduce the whole of the Licensed Material;
 - c. if specified in the Annexure, reproduce part of the Licensed Material;
 - d. if specified in the Annexure, publish the Licensed Material; and
 - e. if specified in the Annexure, communicate the Licensed Material to the public (that is, make the Licensed Material available online or by electronic transmission).

4. The Licence may be exercised only for the limited purpose stated in the Annexure.
5. Unless otherwise stated in the Annexure and subject to paragraph 13, the Licence is perpetual.
6. Unless otherwise stated in the Annexure, the rights granted under the Licence may only be exercised within Australia.
7. If specified in the Annexure, ACARA must approve any work that is created by the Licensee that includes all or part of the Licensed Material prior to the Licensee publishing or communicating the Licensed Material to the public. ACARA's approval will not be unreasonably withheld.

Acknowledgement and notice

8. The Licensee must include the acknowledgement and notice specified in the Annexure in all publications containing all or part of the Licensed Material.

Licence does not affect ownership of Intellectual Property

9. The Licensee agrees that ownership of all intellectual property rights in the Licensed Material remains vested in ACARA.

No warranty by ACARA regarding suitability or fitness for purpose

10. No warranty is given or implied in respect of the suitability or fitness for purpose of the Licensed Material for the purposes specified in the Annexure.
11. Where any statute implies into this letter any term, condition or warranty which would otherwise be excluded by paragraph 10 and that statute prohibits provisions in a contract excluding or modifying the application or exercise of, or liability under, such term, condition or warranty, such term, condition or warranty shall be deemed to be included in this letter, but ACARA's liability for any breach of such term, condition or warranty shall be limited to the extent permitted by law.
12. The Licensee acknowledges that it has not relied on any representations made by ACARA that have not been stated expressly in this letter.

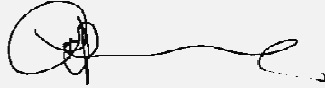
Termination

13. ACARA may, by notice, terminate the Licence with immediate effect.

Further queries

14. If you have any further queries regarding the content of this letter, please contact ACARA at info@acara.edu.au.

Yours sincerely

A handwritten signature in black ink, consisting of a large, stylized capital 'P' followed by a series of loops and a long horizontal stroke ending in a small hook.

Peter Adams
A/General Manager
Assessment
Australian, Curriculum, Assessment and Reporting Authority

Annexure - Permission details

Description of Licensed Material (refer paragraph 1)	Adaptation of text from National Assessment Program – Literacy and Numeracy 2008: Writing – Narrative Marking Guide available at www.naplan.edu.au/verve/_resources/napmarkguide08.pdf <i>Licensed Material attached</i>
Is the Licensee permitted to reproduce the Licensed Material in part? (refer paragraph 3)	Yes
Is the Licensee permitted to publish the Licensed Material? (refer paragraph 3)	Yes
Is the Licensee permitted to communicate the Licensed Material to the public? (refer paragraph 3)	Yes
Purpose (refer paragraphs 4 and 10)	Use the Licensed Material in the PhD entitled "The Theory, Design, Development and Evaluation of the MarkIT Automated Essay Grading System" for non-commercial educational purposes only.
Period of licence (refer paragraph 5)	No change
Territory (refer paragraph 6)	Worldwide
Prior approval for publication etc (refer paragraph 7)	No approval required
Form of acknowledgement and notice (refer paragraph 8)	Adapted from text in <i>National Assessment Program – Literacy and Numeracy 2008: Writing – Narrative Marking Guide</i> . © Australian Curriculum, Assessment and Reporting Authority, 2008. The material is reproduced with the permission of ACARA.

Appendix C Publications

Paper No. 1

Williams, R. (2001) Automated Essay Grading: An Evaluation of Four Conceptual Models in
Kulski, M. & Herrmann, A.(editors) (2001), *New Horizons in University Teaching and
Learning: Responding to Change*, Curtin University of Technology, Perth, Australia

Automated essay grading: An evaluation of four conceptual models

Robert Williams

School of Information Systems, Curtin University of Technology

Automated essay grading has been proposed for over thirty years. Only recently have practical implementations been constructed and tested. This paper describes the theoretical models of four implemented systems described in the literature, and evaluates their strengths and weaknesses. All four models make use of comparisons with one or many model answer documents that have been previously assessed by human markers. One hybrid system that makes use of some linguistic features, combined with document characteristics, is shown to be a practical solution at present. Another system that makes use of primarily linguistics features is also shown to be effective. An implementation that ignores linguistic and document features, and operates on the "bag of words" approach, is then discussed. Finally an approach using text categorisation techniques is considered.

Introduction

Teaching staff around the world are faced with a perpetually recurring problem: how do they minimise the amount of time spent on the relatively monotonous tasks associated with grading their students' essays. With the advent of large student numbers, often counted in thousands in first year common core units, the grading load has become both time-consuming and costly. A system that can automate these tasks is currently just a dream for most staff.

One of the earliest mentions of computer grading of essays in the literature was in an article by Page in which he described Project Essay Grade (PEG). (Page, 1966). Various aspects of students' essays, such as proportion of words on a common word list acting as a proxy for diction, and the proportion of prepositions acting as a proxy for sentence complexity, were measured. A multiple regression technique was then used to predict the human raters' score, based on these measures. We discuss the latest version of PEG later in this article.

Page (1966) made a distinction, which is still relevant today, between grading for content and grading for style:

" 'Content' refers loosely to what the essay says, and 'style' refers to syntax and mechanics and diction and other aspects of the way it is said." (p.240)

This dichotomy gives us the basis for classifying the systems that have been developed. Do they grade primarily for subject matter or for linguistic style? And, do we measure proxies for these dimensions (rating simulation), or do we measure the actual dimensions (master analysis)? Table 1 shows the resulting four categories.

Table 1. Possible Dimensions of Essay Grading (Source: Page, 1966: 240)

	I Content	II Style
A. Rating Simulation	I(A)	II(A)
B. Master Analysis	I(B)	II(B)

There are inherent problems to be overcome if automated grading of text is to become a reality. Student essays addressing a particular topic can theoretically be expressed in possibly thousands of forms, using different combinations of words and sentences. Simply checking for the occurrence of some key words does not allow for a very accurate assessment of the work, nor does it allow for the richness and diversity that English allows for expression of ideas. Many words have thirty to forty entries in a thesaurus, and generally many of them are interchangeable in a particular and given context, so checking for the occurrence of key words is not an acceptable approach.

Conceptual Models for Automated Essay Grading

The first model, Project Essay Grade (PEG), is one of the earliest and longest-living implementations of automated essay grading (Page, 1966). It has been developed by Page and colleagues, and primarily relies on linguistic features of the essay documents.

The second model, E-rater developed by Burstein, Kukich, Wolff, Lu, & Chodorow, (1998) at the Educational Testing Service (ETS) in the USA, has been implemented to the prototype stage for evaluation. This model uses a hybrid approach of combining linguistic features, derived by using Natural Language Processing (NLP) techniques, with other document structure features.

The third model, the LSA model, makes use of Latent Semantic Analysis (LSA) and the "bag of words" approach, and has been developed and evaluated by Landauer, Foltz, & Laham (1998) at the University of Colorado in Boulder. It ignores document linguistic and structure features.

The fourth model, which uses text categorisation techniques, identified in this paper as TCT, has been developed by Larkey (1998) at the University of Massachusetts. It uses a combination of modified key words and linguistic features.

PEG

Description

The idea behind PEG is to help reduce the enormous essay-grading load in large educational testing programs, such as the SAT. When multiple graders are used, problems arise with consistency of grading. A larger number of judges are likely to produce a true rating for an essay.

A sample of the essays to be graded is selected and marked by a number of human judges. Various linguistic features of these essays are then measured. A multiple regression equation is then developed from these measures. This equation is then used along with the appropriate measures from each student essay to be graded, to predict the average score that a human judge would assign.

PEG has its origins in work begun in the 1960's by Page and his colleagues (Page, 1966).

...we coined two explanatory terms: Trins were the intrinsic variables of interest – fluency, diction, grammar, punctuation, and many others. We had no direct measures of these, so began with substitutes: Proxes were approximations, or possible correlates, of these trins. All the computer variables (the actual counts in the essays) were proxes. For example, the trin of fluency was correlated with the prox of the number of words. (Page, 1994, p. 130)

The multiple regression techniques are then used to compute, from the proxes, an equation to predict a score for each student essay. In the research reported in Page (1994), the goal was to identify those variables which would prove effective in predicting human rater's scores. Various software products, including a grammar checker, a program to identify words and sentences, software dictionary, a part-of-speech tagger, and a parser were used to gather data about many proxes.

Evaluation

Details of most of the predictive variables are not given in Page's work. However, amongst the variables found useful in the equation were the fourth root of the number of words, sentence length, and a measure of punctuation. One set of results, based upon a regression equation with twenty-six variables, showed correlations between PEG predicted scores and human rater scores varying between 0.389 and 0.743.

E-rater

Description

E-rater uses a combination of statistical and NLP techniques to extract linguistic features of the essays to be graded. As in all the conceptual models discussed in this paper, E-rater student essays are evaluated against a benchmark set of human graded essays. E-rater has modules that extract essay vocabulary content, discourse structure information and syntactic information. Multiple linear regression techniques are then used to predict a score for the essay, based upon the features extracted. For each new essay question, the system is run to extract characteristic features from human scored essay responses. Fifty seven features of the benchmark essays, based upon six score points in an ETS scoring guide for manual grading, are initially used to build the regression model. Using stepwise regression techniques, the significant predictor variables are determined. The values derived for these variables from the student essays are then substituted into the particular regression equation to obtain the predicted score.

One of the scoring guide criteria is essay syntactic variety. After parsing the essay with an NLP tool, the parse trees are analysed to determine clause or verb types that the essay writer used. Ratios are then calculated for each syntactic type on a per essay and per sentence basis.

Another scoring guide criteria relates to having well-developed arguments in the essay. Discourse analysis techniques are used to examine the essay for discourse units by looking for surface cue words and non-lexical cues. These cues are then used to break the essay up into partitions based upon individual content arguments.

The system also compares the topical content of an essay with those of the reference texts by looking at word usage.

Evaluation

The system has been evaluated by Burstein et al. (1998) and has found that it can achieve a level of agreement with human raters of between 87% and 94%, which is claimed to be comparable with that found amongst human raters. For one test essay question the following predictive feature variables were found to be significant.

1. Argument content score
2. Essay word frequency content score
3. Total argument development words/phrases
4. Total pronouns beginning arguments
5. Total complement clauses beginning arguments
6. Total summary words beginning arguments
7. Total detail words beginning arguments
8. Total rhetorical words developing arguments
9. Subjunctive modal verbs

The LSA model

Description

LSA represents documents and their word contents in a large two dimensional matrix semantic space. Using a matrix algebra technique known as Singular Value Decomposition (SVD), new relationships between words and documents are uncovered, and existing relationships are modified to more accurately represent their true significance.

The words and their contexts are represented by a matrix. Each word being considered for the analysis is represented as a row of a matrix, and the columns of the matrix represent the sentences, paragraphs, or other subdivisions of the contexts in which the words occur. The cells contain the frequencies of the words in each context.

The SVD is then applied to the matrix. SVD breaks the original matrix into three component matrices, that, when matrix multiplied, reproduce the original matrix. Using a reduced dimension of these three matrices in which the word-context associations can be represented, new relationships between words and contexts are induced when reconstructing a close approximation to the original matrix from the reduced dimension component SVD matrices. These new relationships are made manifest, whereas prior to the SVD, they were hidden or latent.

To grade an essay, a matrix for the essay document is built, and then transformed by the SVD technique to approximately reproduce the matrix using the reduced dimensional matrices built for the essay topic domain semantic space. The semantic space typically consists of human graded essays. Vectors are then computed from a student's essay data. The vectors for the essay document, and all the documents in the semantic space are compared, and the mark for the graded essay with the lowest cosine value in relation to the essay to be graded is assigned.

The Intelligent Essay Assessor is a commercial implementation of the LSA approach. Later in this paper we discuss a trial of this system for first year university student essays.

Evaluation

Landauer, et al. (1998), report that LSA has been tried with five scoring methods, each varying the manner in which student essays were compared with sample essays. Primarily this had to do with the way cosines between appropriate vectors were computed. For each method an LSA space was constructed based on domain specific material and the student essays. Foltz (1996) also reports that LSA grading performance is about as reliable as human graders. Landauer (1999) reports another test on GMAT essays where the percentages for adjacent agreement with human graders were between 85%-91%.

The Text Categorisation Technique (TCT)

Description

Larkey (1998) implemented an automated essay grading approach based on text categorisation techniques, text complexity features, and linear regression methods. The Information Retrieval literature discusses techniques for classifying documents as to their appropriateness of content for given document retrieval queries (van Rijsbergen, 1979). Larkey's approach

".. is to train binary classifiers to distinguish 'good' from 'bad' essays, and use the scores output by the classifiers to rank essays and assign grades to them." (Larkey, 1998, p.90)

The technique firstly makes use of Bayesian independent classifiers (Maron, 1961) to assign probabilities to documents estimating the likelihood that they belong to a specified category of documents. The technique relies on an analysis of the occurrence of certain words in the documents. Secondly, a k-nearest neighbour technique is used to find the k essays closest to the student essay, where k is determined through training the system on a sample of human graded essays. The Inquiry retrieval system (Callan, Croft, & Broglio, 1995) was used for this. Finally, eleven text complexity features are used, such as the number of characters in the document, the number of different words in the document, the fourth root of the number of words in the document (see also the discussion on PEG above), and the average sentence length.

Larkey conducted a number of regression trials, using different combinations of components. He also used a number of essay sets, including essays on social studies (soc), where content was the primary interest, and essays on general opinion (G1), where style was the main criteria for assessment. The results presented here are for these two essay sets only.

Evaluation

When all the criteria for assessment were used the proportion of essays graded exactly the same as human graders was 0.60 and scores adjacent (a score one grade on either side) was 1.00. For the general opinion essays the corresponding figures were 0.55 and 0.97. The system performed remarkably well.

Discussion

We are now in a position to characterise these essay grading techniques according to the classification postulated by Page (1966):

PEG focuses on simple linguistic features, focusing on style, and can be categorised as II(A). E-rater focuses on linguistic features and document structures, and is thus performing a Master Analysis of style, and falls in the category II(B). The LSA model focuses on the semantics of the essay, but does so using a Rating Simulation, and therefore falls in the I(A) category. The TCT (soc) experiments focused on content in a rating simulation, while the TCT (G1) test focused on style in a rating simulation. Table 2 summarises these models' classifications.

Table 2. Essay Grading Models' Classifications

	I Content	II Style
A. Rating Simulation	LSA, TCT (soc)	PEG, TCT (G1)
B. Master Analysis		E-rater

Table 3 shows some of the reported performances, in comparison to human graders, of the various models.

Table 3. Comparative performance of models

Model	Measure	Values	Source
PEG	r	0.389-0.743	Page, 1994
E-rater	%	87-94	Burstein, et al, 1998
LSA	%	85-91	Landauer, 1999
TCT (soc)	r	0.69-0.78	Larkey, 1998
TCT (G1)	r	0.69-0.88	Larkey, 1998

To find the amount of total variation explained by a correlation we take its square (PEG performance thus accounts for between 15% and 55% of the variations between PEG and human ratings, and TCT accounts for between 47% and 77%). It appears then, in terms of comparison with human markers, E-rater is best, followed by LSA, TCT, and finally PEG.

Trial of the Intelligent Essay Assessor

A team of researchers in the School of Information Systems at Curtin University of Technology trialled the Intelligent Essay Assessor (IEA) during the first semester of 2001.

In March 2001, students enrolled in the unit Information Systems 100 (IS100) were notified that they could receive bonus marks of up to 5 per cent if they took part in the trial by submitting a two to three page essay based on a question taken from their textbook. These essays, in Microsoft Word format, were submitted via email to a special IS100 email address.

In May, 2001 an honours student in the School converted the essays to a standard format, and added student identification. Two hundred formatted essays were then chosen at random to be graded by three human markers. The average grade for these essays was 64.5. These essays, known as the training set, were sent to the USA to be processed by the IEA to form the semantic (knowledge) space, against which the other essays would be graded.

In June 2001 an additional 327 un-graded essays were sent by email to the USA for IEA grading, and the results were received back one week later. The system produced an average grade of 65.53. The accuracy of the IEA was very good, when compared to the human graded average. Figure 1 shows the distribution of grades produced by the IEA.

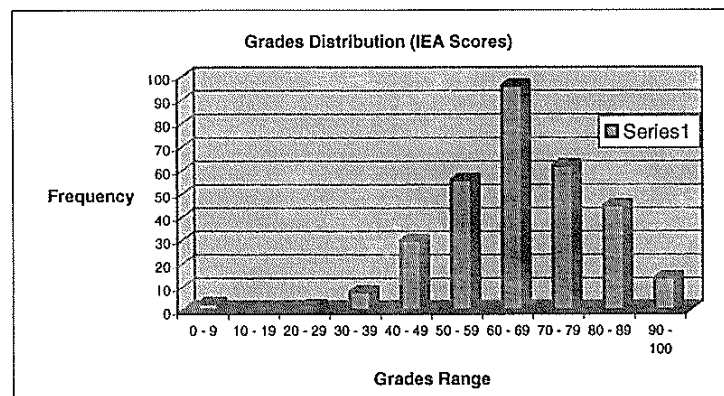


Figure 1. Distribution of grades produced by the IEA

The IEA also detected a number of cases of plagiarism that had escaped the attention of the human graders.

The cost per paper for the automated grading was about A\$30, which is high when compared to human grading costs, but economies of scale apply to the IEA, and this cost could be reduced considerably (to about A\$5) if more papers were graded against the same semantic space.

The researchers felt that the IEA is suitable when very large numbers of essays are to be graded (e.g., 2000), but the effort involved in formatting and human grading 200 essays for the semantic space, and the set-up costs, are too great when only a few hundred essays are to be graded. The researchers were impressed by the ability of the IEA to detect plagiarism amongst the essays submitted by the students.

Conclusion

Automated essay grading is now ready to advance from the research laboratory to the real world educational environment. Current prototype systems, which grade for content, style, or both, can perform equally well as human graders. Prototype systems only need minor enhancements to move into educational systems worldwide. However, they cannot at present deal with tabular and graphical content in essays. The administrative resources needed to support these systems are quite substantial. Human judges are still needed to prepare model answers, or to grade samples of student essays before the computer systems complete the task. Students also need suitable computer facilities to generate their essays in machine-readable form. It is likely that commercial essay grading products will appear in the next ten years, and help ease the grading workload for teachers in a variety of disciplines.

References

- Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998) *Enriching automated essay scoring using discourse marking*. Proceedings of the workshop on discourse relations and discourse markers. Annual meeting of the association of computational linguistics. Montreal, Canada.

- Callan, J. P., Croft, W. B., & Broglio, J. (1995) TREC and TIPSTER experiments with INQUERY. *Information Processing and Management*, 327-343.
- Foltz, P. W. (1996) Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers*, 28, 197-202.
- Landauer, T. K., Foltz, P. W., and Laham, D. (1998) An introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.
- Landauer, T. K. (1999) Email communication with author. 8th June.
- Larkey, L. S. (1998) Automatic essay grading using text categorization techniques. *Proceedings of the Twenty First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 90-95.
- Maron, M. E. (1961) Automatic indexing: An experimental inquiry. *Journal of the Association for Computing Machinery*, 8, 404-417.
- Page, E. B. (1966) The imminence of grading essays by computer. *Phi Delta Kappan*, 238-243.
- Page, E. B. (1994) Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, 62, 127-142.
- Page, E.B. & Petersen, N.S. (1995) The computer moves into essay grading. *Phi Delta Kappan*, 561-565.
- Perelman-Hall, D. (1992) A natural solution. *Byte*, 17(2), 237-244.
- Salton, G. (1988) *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, Massachusetts: Addison-Wesley,
- van Rijsbergen, C. J. (1979) *Information Retrieval*. (2nd ed.), London: Butterworths.

Paper No. 2

Williams, R., and Dreher, H. (2004). Automatically Grading Essays with MarkIT. *Journal of Issues in Informing Science and Information Technology*, Vol. 1, pp. 693-700.

Automatically Grading Essays with Markit[®]

Robert Williams and Heinz Dreher

Curtin University of Technology, Perth, WA, Australia

williamsr@cbs.curtin.edu.au dreherh@cbs.curtin.edu.au

Abstract

Markit[®] is an Automated Essay Grading (AEG) system capable of running on typical desktop PC platforms. Its performance compares favourably with human graders and with commercially available systems. A distinct advantage of Markit over existing commercial systems is that it requires only one model answer against which the student essays are compared. In this paper we report on a trial of Markit with second year law students' essays.

Keywords: Automated Essay Grading (AEG); Natural Language Processing (NLP); semantics; electronic thesaurus.

Introduction

At IS2002 in Cork, Ireland (Palmer et al. 2002), we reported on a trial of a commercially available computer grading system as used to automatically grade first year university student essays. The results were encouraging, but we felt confident that certain perceived limitations could be overcome by applying ourselves to building our own system. Since then we have developed our own prototype and are investigating its performance with a wide range of subject areas and year levels. We have named our system Markit[®].

In order to automate the grading of essays some method of capturing the meaning of the words, sentences and paragraphs must be found. Representing the semantics of a text document for computational uses is one method but is problematic. How can we formally code the meanings of words, phrases, sentences, paragraphs and so on, so that useful computational work can be done robustly, effectively and efficiently? Such semantic representation and computational work has been applied to the problem of essay grading in an endeavour to overcome some of the limitations we discussed with alternate approaches adopted by existing essay grading systems, but it may also be applied to related problems of text understanding, question answering, and qualitative feedback on assignments, for example.

In this article we present our work in developing the Markit system, now in prototype form and being used for Automated Essay Grading (AEG) and related applications.

Problems Identified from Previous Experience

The system we previously trialed, required us to manually grade 200 essays which were used to

Material published as part of this journal, either on-line or in print, is copyrighted by Informing Science. Permission to make digital or paper copy of part or all of these works for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage AND that copies 1) bear this notice in full and 2) give the full citation on the first page. It is permissible to abstract these works so long as credit is given. To copy in all other cases or to republish or to post on a server or to redistribute to lists requires specific permission from the publisher at Publisher@InformingScience.org

build a reference database against which candidate essays would be benchmarked and an appropriate score assigned. Clearly, if it were possible to use just one model answer to produce a reliable automated scoring system, it would be feasible to employ the system in the grading of tens and hun-

dreds rather than thousands of essays. Such a system would be more widely applicable. Markit uses just one model answer. The content of the model answer is determined by the instructor, and thus he/she has complete control over the knowledge against which the student essays are to be assessed. The breadth and depth of the knowledge required for very good answers can thus be determined by the instructor.

The cost of using the previous system was prohibitive for all but very large (in excess of 2000) essay numbers. Comparing costs can be an arbitrary exercise, and whilst the previously trialed system was approximately double that of manual grading, costs associated with Markit will need to be kept well below the cost of human grading if it is to come into popular usage. At this stage, we may be flexible in assigning costs to essay grading via Markit, preferring to acquire some experience with its use in a wide variety of circumstances. There are real costs however in doing any grading work, and these must be covered on a case by case basis. The costs which may be associated with product development will need to be recouped over a longer term when we are satisfied that Markit is performing at its optimum.

An important aspect of assessment relates to elapsed time taken to complete a job. Invariably, the requirement is for graders, human or otherwise, to produce results within a few days. We found that the elapsed time related to job submission and results provision can be longish, and figured that the availability of an in-house system would permit better control over this factor. Naturally, there would be many ways to control the elapsed-time variable, but having a system at one's fingertips is reassuring at the very least, and permits re-runs, and other experimentation to occur. From this standpoint, Markit suits our purposes very well indeed; we have control of the entire process from assignment and model answer submission through to results notification, meaning that security for example can be managed directly and effectively.

In our previous work we noted another serious limitation to an essay grading system – it can only grade by making a comparison with a given set of subject material. The model answer contains only a set body of knowledge and would grade the student on the part of that knowledge the student was able to demonstrate. Under such circumstances a 'brilliant' answer or essay, for example which drew on associations with material not part of the model answer, would score poorly. This problem, whilst recognized, has not been overcome in Markit's design. We will need to understand that Markit's grades are recommendations to human examiners and should be reviewed for appropriateness much as human scored essays required moderation and review in a selection of cases.

Probably the most advantageous aspect of Markit is its reliance on generally available technology as compared with specialized computing platforms needed to run a previously trialed system. Markit will operate on a standard Windows PC, and it will be developed in such a way that some of the system's components can be shrink-wrapped for widespread distribution and local use. This has been made possible due to the computational algorithm at the heart of Markit's design.

The Markit© System

Some readers will naturally be eager to discover the precise design of Markit's algorithms, but will also understand the proprietary nature of potentially commercially viable systems and thus the confidential nature of those designs. We are able however to characterize Markit's general design, although we look forward to satisfying readers' curiosity by performance data as we progress and expand our use across a wide variety of cases.

The Markit system relies on building a propriety representation of the knowledge contained in the model answer. A student essay is processed using a combination of NLP (Natural Language Processing) techniques to build the corresponding propriety knowledge representation. Pattern matching techniques are then employed to ascertain the proportion of the model answer knowl-

edge that is present in the student answer, and a grade assigned accordingly. An electronic version of Roget's Thesaurus is used to extract lexical information for the building of the document knowledge representation.

The technique allows a formal representation of free unseen text to be quickly and robustly built for further analysis by the Markit system. The approach used has a need for a semantic representation that does not need substantial hand coding of knowledge structures prior to use, and that can deal with unlimited unseen text. Many Natural Language Processing (NLP) systems use some kind of a parser to initially extract the syntax of sentences in a document as an initial step prior to further processing. Semantic analysis then follows. The use of Context Free Phrase Structure Grammar (CFPSG) parsers is commonly suggested in the literature. However CFPSG parsing cannot be used in all but simple toy domains. The reason for this is that free unseen text is very hard to parse, because the set of grammar rules required is very large, and the time taken to evaluate every possible parse tree generated is too great for a practical system. So while CFPSG parsing has been tried with the prototype system described in this article, it has been abandoned in favour of using "Chunking" to determine the phrases and clauses used for further processing. "Chunking" enables one to use grammar heuristics to derive noun phrases and verb clauses very quickly from unseen text. The problem of unrealistic parsing time is thus eliminated.

Markit's grading system is capable of producing perfect scores when grading a document against itself, the desirability of which can be appreciated.

Markit is at the prototype stage, and is undergoing further development. The system currently has 8 subsystems written in C++, Java and Visual Basic for Applications.

The extraction of information from Roget's Thesaurus (Roget, 1991) is slow, due to the fact that approximately 500 pages of a Microsoft Word document have to be scanned for each word in a sentence, using Visual Basic for Applications code. This process can take up to about 10 minutes for a 40 word sentence and clearly needs to be modified to access a database version of Roget's Thesaurus or equivalent. We are currently in negotiations with a publisher to have research access to a commercially available electronic dictionary/thesaurus. We will incorporate this in a database table, facilitating direct file access to each word in the thesaurus. Processing time will then be reduced to matter of seconds.

System performance otherwise is very good, with the non-Roget's Thesaurus related work taking only a few seconds for a 2 page document on a 1.9 Ghz Pentium 4 processor.

Markit Performance - the Law 252 Trial

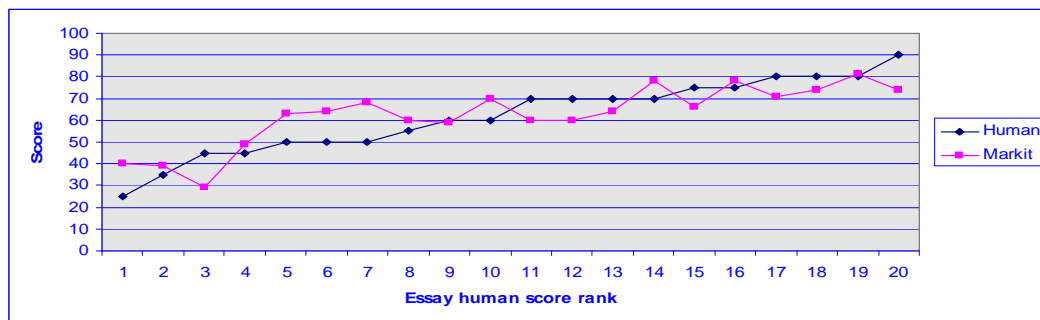
Markit was trialed with student essays from a second year university law unit taught at the Curtin Business School (Willesee, 2003).

A lecturer in a School of Business Law kindly volunteered to assist by providing 66 essays from his unit Law 252 for Markit to assess. He also provided a model answer. This model answer was then processed to produce the model answer numerical summary. A selection of 20 student essays were then processed and graded against the model answer. Comparisons were made between the human scores and the automated scores from Markit. The results are shown in Table 1 and Figure 1.

Table 1 - Human and Markit scores for Law 252 essays

Essay	Human	Markit	Absolute Difference	Human vs Markit
1	25	40	15	0.79
2	35	39	4	
3	45	29	16	
4	45	49	4	
5	50	63	13	
6	50	64	14	
7	50	68	18	
8	55	60	5	
9	60	59	1	
10	60	70	10	
11	70	60	10	
12	70	60	10	
13	70	64	6	
14	70	78	8	
15	75	66	9	
16	75	78	3	
17	80	71	9	
18	80	74	6	
19	80	81	1	
20	90	74	16	
Average	61.75	62.35	8.90	

From Table 1 (essay 1) it can be seen that a student essay was graded 25/100 by a human and 40 by Markit, resulting in a difference of 15 percentage points between the two. Now this student would presumably be very happy with such a score since it is higher than that of the Human grader. However the student author of essay 3 suffers from about the same error in absolute terms, but surely would be grossly dissatisfied with the failing grade. These extreme cases will need to be investigated with a view to understanding the reason. We have found in one much more extreme case that the error was due to the human grader rather than Markit. Obviously,

**Figure 1 - Human versus Markit scores in ascending order of human scores for Law 252 essays**

many more studies need to be conducted across time, subject matter, year levels, and so on, before we can appreciate the true worth of Markit's contribution. So, whilst the average differences between Human grades and Markit is acceptable, it is the large individual differences, particularly where the Markit grade is lower which need investigation and analysis.

Note the difference in the average marks is 0.60 %, and this is not significant with a p-value of 0.80 for a 2 tailed test of significance. The average error is 8.90%, and the Pearson correlation between the human and computer scores is 0.79, significant at the 1% level with a 2 tail test. This correlation is computed on the two scores for the same essay.

When we rearrange the data in ascending order of Markit scores, Figure 2 shows the trend. It appears that Markit is assigning scores using a moving average of the three neighbouring human scores. This is a highly desirable outcome, indicating that it is capturing the essence of the human grader's criteria, but of course Markit does not have access to the human scores. This characteristic has not yet been analysed statistically.

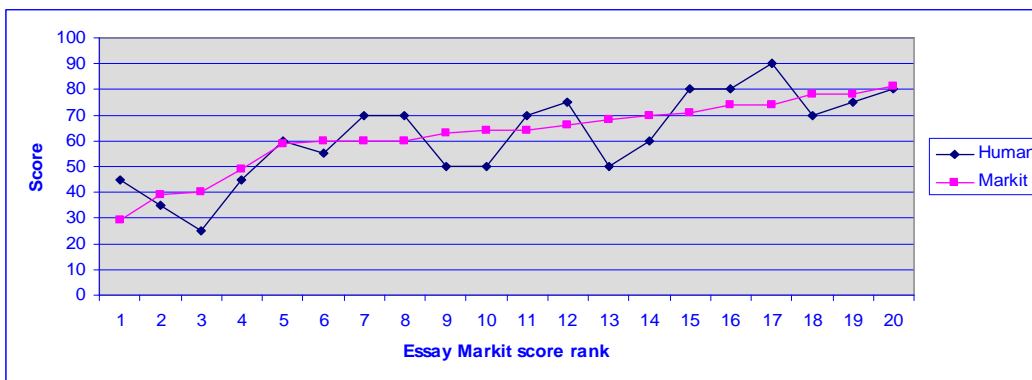


Figure 2 - Human versus Markit scores in ascending order of Markit scores for Law 252 essays

Markit Performance – compared to the IEA

As already mentioned, in 2001 we conducted a trial of a commercially available automated essay grading system (the IEA – Intelligent Essay Assessor, see Landauer et al. 1998) using essays from a first year Curtin University unit, Information Systems 100.

Nine of these essays were also graded by Markit. The top graded essay by the IEA gained 99%. This essay was then used as the model answer against which the others were compared using Markit. Table 2 and Figure 3 below show the results.

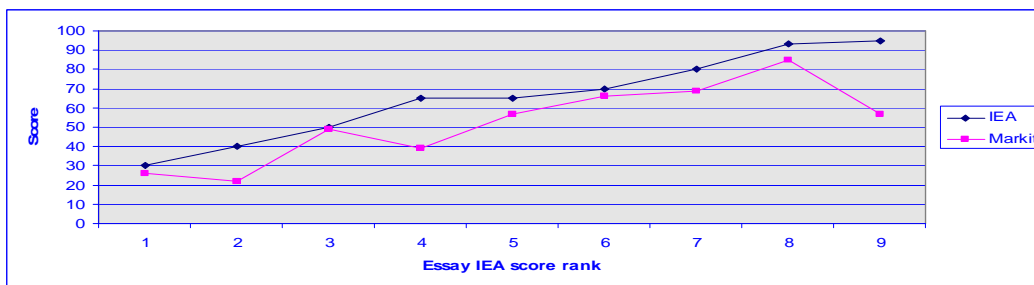


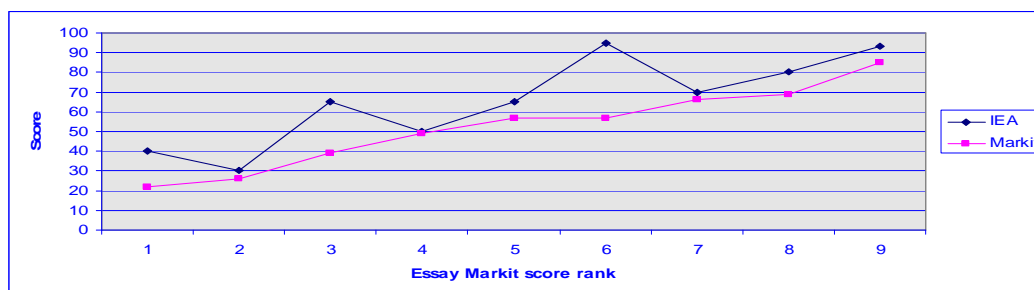
Figure 3 - IEA versus Markit scores in ascending order of IEA scores for IS 100 essays

Table 2 - IEA and Markit scores for IS 100 essays

Essay	IEA	Markit	Absolute Difference	IEA vs Markit
1	30	26	4	0.84
2	40	22	18	
3	50	49	1	
4	65	39	26	
5	65	57	8	
6	70	66	4	
7	80	69	11	
8	93	85	8	
9	95	57	38	
Average	65.33	52.22	13.11	

Note the difference in the average marks is 13.11 %, and this is significant with a p-value of 0.01 for a 2 tailed test of significance. The average error is 13.11%, and the Pearson correlation between the human and computer scores is 0.84, significant at the 1% level with a 2 tail test. This correlation is computed on the two scores for the same essay.

When we rearrange the data in ascending order of Markit scores, Figure 4 shows the trend. In this case, Markit appears to be assigning scores on a downward shifted moving average of the three neighbouring IEA scores, again not analysed statistically. This is not an ideal situation, but an adjustment factor could be built into Markit to shift the scores up. Further testing of Markit will determine the extra tuning parameters we may need to add to the scoring algorithm. Again, Markit does not have access to the IEA scores.

**Figure 4 - IEA versus Markit scores in ascending order of Markit scores for IS 100 essays**

Markit grades well when compared to the human marker. It tracks the human scores well, with an acceptable error rate. However it appears that we have tuned Markit to the Law 252 essays, as it scores lower than the IEA on the IS 100 essays, which has an average error of 13.11%. The Pearson correlations in both cases are acceptable, although we would like to see them at about 0.90. The IEA essays were not directly graded against a human marker, and the lower scores from Markit could be related to this.

The result of Markit's automated grading, when compared to human markers, is at the high end of published results for other AEG systems (Williams, 2001). Dessus et al., (2000) report that the highest correlations are found between human graders and Latent Semantic Analysis (LSA) based

techniques and are 0.80 and 0.86. In a trial of Markit we obtained correlations of 0.79 between the human marker and Markit, and, on a different set of essays, 0.84 between the Intelligent Essay Assessor grading and Markit grading.

Concluding Observations

The semantic representation as mentioned above lends itself to speedy processing by the grading subsystem. Comparisons between documents can then be made by looking for similar content, even if the documents use completely different wording. The programming of such content matching algorithms is relatively straightforward as a result of thesaurus based numerical array structure representations as used by Markit for comparing student essays against model answers.

Markit's performance is equally as good as other systems documented in the literature and yet the performance is achieved with minimal human grader input – only one model answer is required in comparison with some other systems which require several hundred human graded essays.

References

- Dessus, P., Lemaire, B., & Vernier, A. (2000). Free-text assessment in a virtual campus. *Proceedings CAPS'2000*. Paris : Europa, 13 –14 Dec.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis, *Discourse Processes*, 25, 259-284.
- Palmer, J., Williams, R., & Dreher, H. (2002). Automated essay grading system applied to a first year university subject – How can we do it better? *2002 Informing Science & IT Education Joint Conference: InSITE*, University College Cork, Ireland.
- Roget, P. M. (1991). *Roget's thesaurus*. Project Gutenberg, <http://promo.net/pg/>
- Williams, R. F. (2001). Automated essay grading: An evaluation of four conceptual models. In M. Kulski & A. Herrmann (Eds.), *New horizons in university teaching and learning: Responding to change*. Curtin University of Technology, Perth, Western Australia
- Willesee, W. (2003). Can a computer grade an Internet uploaded law essay? *5th Conference on Computerisation of Law via the Internet*, Sydney, Australia. November.

Biography

Robert Williams has over 25 year's experience in the Information Systems industry, as a practitioner, researcher and lecturer. He currently is a lecturer in the School of Information Systems at Curtin University of Technology in Perth, Western Australia. He has extensive experience in systems analysis and design, and programming, on a variety of mainframe, mini and personal computers, and a variety of operating systems and programming languages. Applications he has worked with include mathematical, statistical, bridge and road engineering, financial, corporate resource allocation, business simulation and educational systems. He has published a number of articles on system users' personalities and satisfaction, decision support systems, and automated essay grading systems. In 2001 he led a team of researchers in the School of Information Systems at Curtin University of Technology which conducted what is believed to be the first trial in Australia of an Automated Essay Grading system. Robert holds a Bachelor of Arts degree with double majors in Mathematics and Economics from the University of Western Australia, a Graduate Diploma in Computing from the Western Australian Institute of Technology, and a Master of Information Systems degree from Curtin University of Technology.

Heinz Dreher has been working in the Information Technology Systems domain for 33 years. His first position was as computer programmer. This was followed with a move into the tertiary education sector in 1972 as senior tutor in Electronic Data Processing (EDP). Currently he is on

sabbatical at the Institute of Interactive Computer Multimedia at the University of Technology Graz, Austria. His substantive position is in the School of Information Systems at Curtin University of Technology, Perth, Western Australia. Dr Dreher has expertise in Hypertext/Hypermedia systems and textual-knowledge-based systems, Computer Supported Co-operative Work (CSCW), Computer Mediated Communications (CMC), Project Management, Prototyping systems, Human Problem Solving Strategies, Decision Support Technologies, Knowledge Management, WWW and Electronic Commerce applications development and technologies, and Information Systems Research Methods. The Hypertext Research Laboratory, whose aim is to facilitate the application of hypertext-based technology in academe, business and in the wider community, was founded by him in late 1989.

Williams, R. and Dreher, H. (2005). Telecommunications Use in Education to Provide Interactive Visual Feedback on Automatically Graded Essays. *Proceedings of International Telecommunications Society Africa-Asia-Australasia Regional Conference*, Perth, Australia.

This paper is unable to be reproduced here due to proprietary reasons.

Paper No. 4

Williams, R., and Dreher, H. (2005). Formative Assessment Visual Feedback in Computer Graded Essays. *Journal of Issues in Informing Science and Information Technology*, Vol. 2, pp. 23-32.

Formative Assessment Visual Feedback in Computer Graded Essays

Robert Williams and Heinz Dreher
Curtin University of Technology, Perth, Western Australia

bob.williams@cbs.curtin.edu.au h.dreher@curtin.edu.au

Abstract

In this paper we discuss a simple but comprehensive form of feedback to essay authors, based on a thesaurus and computer graphics, which enables the essay authors to see where essay content is inadequate in terms of the discussion of the essay topic. Concepts which are inadequately covered are displayed for the information of the author so that the essay can be improved. The feedback is automatically produced by the MarkIT Automated Essay Grading system, being developed by Curtin University researchers.

Keywords: AEG, Automated Essay Grading, visualisation, automated assignment assessment, formative assessment, graphical representation.

Background

The motivation for developing computer supported techniques to assess or grade free text assignments or essays is rather obvious - increased speed, efficiency and consistency, and thus reduced costs and an amelioration of the onerous nature of (humans) marking large volumes of essays in a short time. Of course, this assumes effectiveness, reliability and user (student and teacher) acceptance of 'computer as assessor'. These three aspects have been reported on in the work of Williams & Dreher (2004) for example.

Automated Essay Grading (AEG) is an emerging phenomenon widely documented in the literature (Shermis & Burstein, 2003; Valenti, Neri & Cucchiarelli, 2003; Williams, 2001; Williams & Dreher, 2004). Many of the current AEG systems claim to produce various kinds of feedback regarding the knowledge deficit or other problems in the essays enabling the essay authors to learn, improve, and correct the errors for future submissions. However, much of the feedback is generic in form, for example "this section is inadequate" or "this section needs improvement". This sort of feedback is not very helpful to the learner, and if the truth be known, it is often provided as a justification for the mark, so that when a student queries the grade given, the assessor can offer some further 'soothing' words at least not inconsistent with the original feedback. Of course, the type of evaluation we are concerned with here is formative, and we appreciate that the case of summative evaluation needs to be treated separately – our interest is in the former.

Purpose of Assessment

In our work on grading and assessment we take the view that incremental improvement is an important goal for the learner and the teacher. This implies that when students are given assignments it is the teacher's role to evaluate the work against the stated assignment assessment criteria and provide the student with a grade and some reasons which explain why the particular grade was awarded. An example of such a scheme can be seen in Figure 1 for a course dealing with JavaScript programming and website development.

criterion	mark
1) Features	/10
<i>Minimum of 10 features to be listed</i>	
2) Functionality	/10
<i>Implemented features must be purposeful and function correctly</i>	
3) Navigation	/10
<i>Website must be navigable with navigation support</i>	
4) Usability	/10
<i>Website must have good usability</i>	
5) JavaScript code & explanation	/50
<i>5 functions implemented from the suggested list – mark out of 10 for each of 5 functions (5 for code + 5 for explanation)</i>	
6) Innovative aspects	/10
<i>Anything new, different, & exciting; Zero is the default mark; Nominate your candidate feature</i>	
Total score	/100

Figure 1: example of assignment assessment criteria for an interactive website

Note: a third column headed “assessor’s comments” is used to provide constructive feedback

Source: from the authors’ coursework teaching

Naturally, the criteria given in Figure 1 must be distributed with the assignment specification; else the students’ would have no goal. The assessment task for such assignments involves considering the assignment from the viewpoint of each of the six criteria and making some judgment and generating relevant comments.

Assignment tasks which can conveniently be subdivided into chunks, an extreme example being Multiple-Choice or True-False Tests, lend themselves to computer scoring. However the more essay-like the assignment task the greater the challenge for automated or semi-automated assessment. Nevertheless, there is a growing body of literature in the field of AEG – see below.

In an interesting case of formative evaluation in a course with well in excess of one hundred students, and the flexibility for the students to choose from a variety of topics or themes (Dreher, Scerbakov & Helic, 2004), the authors claim good support provided by the Learning Management System (WBT-Master), which permits individual and relevant formative evaluation comments to be efficiently generated. Figure 2 is a screenshot of an essay assignment being assessed and commented upon.

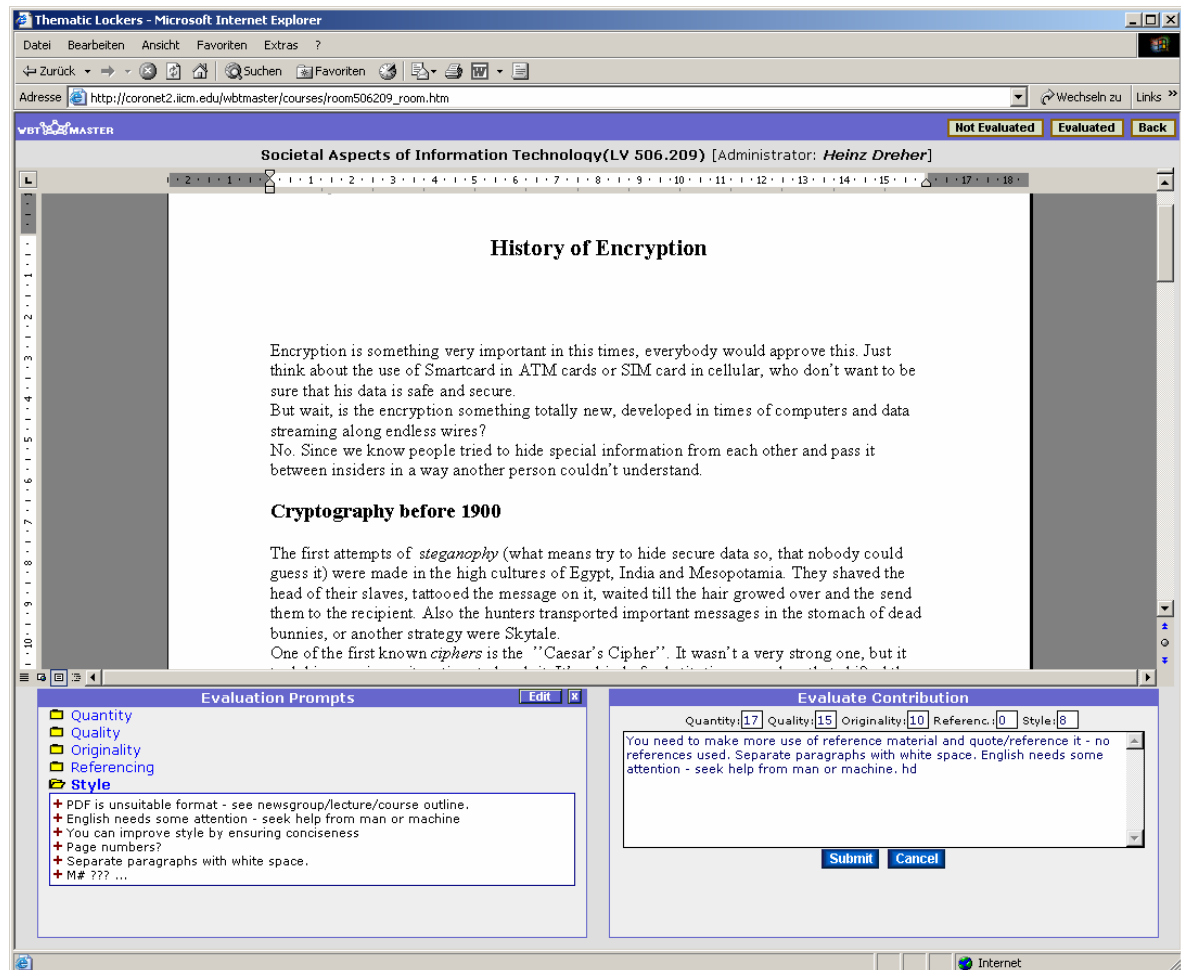


Figure 2 – semi-automated assessment feedback provision for essay assignments

Source: Dreher, Scerbakov & Helic (2004) – reproduced with permission

It should be clear that evaluating assignments and providing feedback to students for the purposes of improvement is on the one hand good education practice, and on the other is very 'expensive'. As we have been developing our AEG system (MarkIT) we have had a unique opportunity to ponder on the provision of meaningful, relevant, consistent feedback which students can use to reflect on their own performance in essay writing.

We now present a short section on the state of the art of AEG, making particular note of the nature and extent of feedback which is provided by these systems, and then take the opportunity to explain how our AEG system has been engineered in terms of feedback provision.

Automated Essay Grading Systems and Feedback Provision

AEG systems are now emerging from the research laboratories into primary, secondary and tertiary education systems around the world (Shermis & Burstein, 2003; Valenti, Neri & Cucchiarelli, 2003; Williams & Dreher, 2004). In the four systems mentioned below, which can be considered as representative of the various approaches to AEG, we consider the level and the form of feedback provided to students. We note that the emphasis is on the grade and not on feedback which may be used to guide improvement and thus further learning. Formative evaluation, including that of content, is considered to be an important aspect of assessment and hence we have worked at including such functionality in MarkIT.

One of the earliest systems for computer grading of essays in the literature was reported in an article by Page in which he described Project Essay Grade (PEG) (Page, 1966). With the rapid advancement in computing power and text processing technologies since the 1960's, more powerful essay grading systems have emerged, and we now discuss the most serious contenders in the field.

PEG

PEG has its origins in work begun in the 1960's by Page and his colleagues (Page, 1966). The idea behind PEG is to help reduce the enormous essay grading load in large educational testing programs, such as the Scholastic Aptitude Test (SAT) (College Board, 2002). When multiple graders are used, problems arise with consistency of grading. A larger number of judges are likely to produce a true rating for an essay. A sample of the essays to be graded is selected and marked by a number of human judges. Various linguistic features of these essays are then measured. A multiple regression equation is then developed from these measures. This equation is then used, along with the appropriate measures from each student essay to be graded, to predict the average score that a human judge would assign. It appears that the main form of this feedback is an essay score, which indicates the level achieved by the student who wrote the essay:

“The feedback provided suggests whether or not students are on a trajectory to take college-level coursework and what remedial options the district offers for those who are not on that trajectory” (Shermis, Mzumara, Olson, & Harrington, 2001, p 248).

E-rater

E-rater uses a combination of statistical and Natural Language Processing (NLP) techniques to extract linguistic features of the essays to be graded. As in all the conceptual models discussed in this paper, E-rater student essays are evaluated against a benchmark set of human graded essays. E-rater has modules that extract essay vocabulary content, discourse structure information and syntactic information. Multiple linear regression techniques are then used to predict a score for the essay, based upon the features extracted. For each new essay question, the system is run to extract characteristic features from human scored essay responses. Fifty seven features of the benchmark essays, based upon six score points in an Educational Testing Services (ETS) scoring guide for manual grading, are initially used to build the regression model. Using stepwise regression techniques, the significant predictor variables are determined. The values derived for these variables from the student essays are then substituted into the particular regression equation to obtain the predicted score. One of the scoring guide criteria is essay syntactic variety. After parsing the essay with an NLP tool, the parse trees are analysed to determine clause or verb types that the essay writer used. Ratios are then calculated for each syntactic type on a per essay and per sentence basis. Another scoring guide criterion relates to having well-developed arguments in the essay. Discourse analysis techniques are used to examine the essay for discourse units by looking

for surface cue words and non-lexical cues. These cues are then used to break the essay up into partitions based upon individual content arguments. The system also compares the topical content of an essay with those of the reference texts by looking at word usage. Given that a detailed analysis of the essay is done it is possible to provide some detailed feedback. A commercial implementation of E-rater is known as Criterion. Criterion feedback gives details of errors in grammar, usage, and mechanics. Other comments about the essay style are also provided. Criterion also provides feedback relating to the essay background, thesis, main ideas, supporting ideas and conclusion (Attali & Burstein, 2004).

IEA

The Intelligent Essay Assessor (IEA) is a Latent Semantic Analysis (LSA) based system. LSA represents documents and their word contents in a large two dimensional matrix semantic space. Using a matrix algebra technique known as Singular Value Decomposition (SVD), new relationships between words and documents are uncovered, and existing relationships are modified to more accurately represent their true significance. The words and their contexts are represented by a matrix. Each word being considered for the analysis is represented as a row of a matrix, and the columns of the matrix represent the sentences, paragraphs, or other subdivisions of the contexts in which the words occur. The cells contain the frequencies of the words in each context. The SVD is then applied to the matrix. SVD breaks the original matrix into three component matrices that, when matrix multiplied, reproduce the original matrix. Using a reduced dimension of these three matrices in which the word-context associations can be represented, new relationships between words and contexts are induced when reconstructing a close approximation to the original matrix from the reduced dimension component SVD matrices. These new relationships are made manifest, whereas prior to the SVD, they were hidden or latent. Landauer, Foltz & Laham (1998) developed the Intelligent Essay Assessor, using the LSA model. To grade an essay, a matrix for the essay document is built, and then transformed by the SVD technique to approximately reproduce the matrix using the reduced dimensional matrices built for the essay topic domain semantic space. The semantic space typically consists of human graded essays. Vectors are then computed from a student's essay data. The vectors for the essay document, and all the documents in the semantic space are compared, and the mark for the graded essay with the lowest cosine value in relation to the essay to be graded is assigned. Such techniques would presumably permit detailed feedback provision - the system gives an estimated grade for the essay, and also details of subtopics that the student did not cover in the essay. (Foltz, Laham, & Landauer, 1999).

TCT

Larkey (1998) implemented an AEG approach based on text categorization techniques (TCT), text complexity features, and linear regression methods. The Information Retrieval literature discusses techniques for classifying documents as to their appropriateness of content for given document retrieval queries (van Rijsbergen, 1979). Larkey's approach

“.. is to train binary classifiers to distinguish “good” from “bad” essays, and use the scores output by the classifiers to rank essays and assign grades to them.” (Larkey, 1998, p90)

The technique firstly makes use of Bayesian independent classifiers (Maron, 1961) to assign probabilities to documents estimating the likelihood that they belong to a specified category of documents. The technique relies on an analysis of the occurrence of certain words in the documents. Secondly, a k-nearest neighbour technique is used to find the k essays closest to the student essay, where k is determined through training the system on a sample of human graded essays. The Inquiry retrieval system (Callan, Croft, & Broglio, 1995) was used for this. Finally, eleven text complexity features are used, such as the number of characters in the document, the

number of different words in the document, the fourth root of the number of words in the document, and the average sentence length. Larkey conducted a number of regression trials, using different combinations of components. She also used a number of essay sets, including essays on Social Studies, where content was the primary interest, and essays on general opinion, where style was the main criterion for assessment. This system appears to only provide a discrete grade for each essay processed (Larkey, 1998).

The MarkIT Automated Essay Grading System

MarkIT is an AEG system that uses propriety technology based on NLP techniques, which has at its core an electronic thesaurus (Williams & Dreher, 2004). As with some other AEG systems, 50-200 human graded essays are used to build a scoring algorithm using multiple linear regression. Better performance is obtained if multiple humans grade the same essays and the scores averaged. An instructor prepares an electronic model answer on the essay topic. Typically this is done with reference to the assignment objectives and assessment criteria. In practice the model answer is often represented as 'the best' of the human graded essays, as instructors may not have developed as clear a formulation of 'good' answers as would be desirable. Students electronically submit their essays on the topic, via the web. The model answer is processed by the system to build up a propriety representation of the meaning of the content of the essay. Student answers are processed in the same manner. The student answers are then processed to ascertain how much of the model answer's content is contained in them. Grades are assigned accordingly.

The MarkIT system relies on building a propriety representation of the knowledge contained in the model answer. A student essay is processed using a combination of NLP techniques to build the corresponding propriety knowledge representation. Pattern matching techniques are then employed to ascertain the proportion of the model answer knowledge that is present in the student answer, and a grade assigned accordingly. An electronic version of a thesaurus is used to extract lexical information for the building of the document knowledge representation.

The technique allows a formal representation of free unseen text to be quickly and robustly built for further analysis by the MarkIT system. The approach used has a need for a semantic representation that does not need substantial hand coding of knowledge structures prior to use, and that can deal with unlimited unseen text. Many NLP systems use some kind of a parser to initially extract the syntax of sentences in a document as an initial step prior to further processing. Semantic analysis then follows. MarkIT uses a specially designed chunking algorithm to perform preliminary processing to extract noun phrases and verb clauses contained in essay sentences.

First experiences show good performance. Experiments have been conducted with a number of 1st year Information Systems student essays, and 2nd year Law student essays, both at university level, and also year 8 secondary school English essays. These essays were prepared by students using a word processor, and comprised some 300 to 500 words, or about one page of text. Expert human graders created the "Human" scores in the usual way by applying the model answer criteria to the essays presented to grading. The computer scoring was a rather simple process of compiling all student answers into text files and submitting them to the computer algorithm. Our technology takes less than 5 seconds per essay to deal with the types of inputs described above. Feeding the model answer which is derived from the course content to the computer is a slightly more involved task.

MarkIT Results for 20 Law Essays

The graph in Figure 3 - Human vs Computer-based scores in ascending order of Human scores represents results for a sample of 20 law essays (horizontal axis) in which the maximum possible assessment was 30 (vertical axis) and shows the comparison between expert human and computer

assessments. The data is (arbitrarily) ordered by increasing computer score. Assignment 1 is assessed by the human at 2 and by the computer at 10 (leftmost data item). Assignment 10 is assessed at 21 by both human and computer, whereas assignment number 20 (rightmost data point) is assessed by the human at 27, and by the computer at 32 – yes, we omitted to inform the computer about the maximum mark on this run! As can be seen the computer tracks the human reasonably well, but further scoring algorithm refinement is indicated. The correlation between the human and computer scores is 0.72.

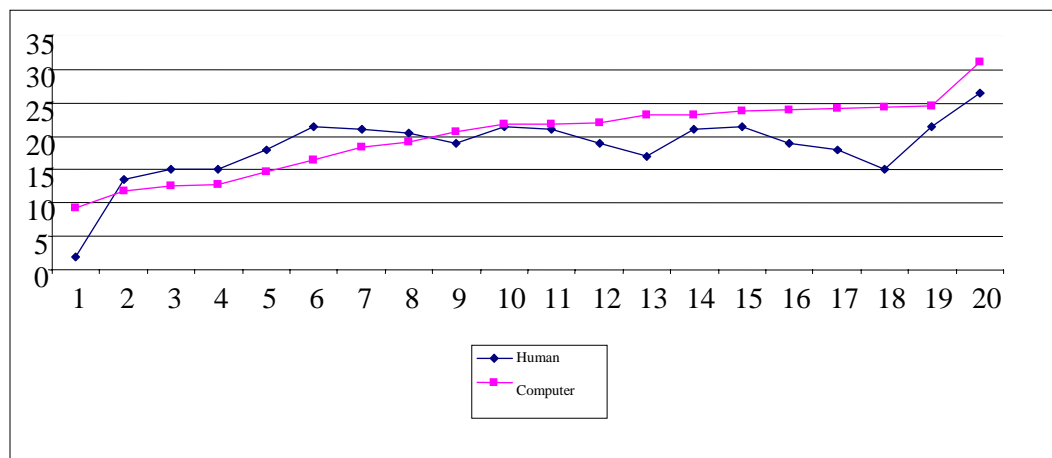


Figure 3 - Human vs Computer-based scores in ascending order of Human scores

Graphical Feedback

In Figure 4 – Concept frequencies: student answer and course content, we have presented another example of MarkIT output. In this case we have a graph showing the ‘concepts’ associated with both the model answer and the student answer. Naturally, the better the correspondence between the concept representation in both, the better the score. If we focus on the tallest bar (Concept_Number 31) we see that the student answer (dark bar) contains a concept_frequency of 6 (vertical axis) where the model answer called for no discussion on this topic or concept. We say the student has introduced irrelevancies into the answer; or perhaps this is what can be termed an error on the student’s part. Concept_Number 26 has a better match between model and student answer, indicating the student has learned relevant material. There are three cases where the model answer concepts are not matched by a student contribution (3, 28, 30) – this we would call “ignorance” or a deficit in knowledge. Such visual feedback is rather informative to student and teacher alike. It is intended to further develop such visual feedback into a dynamic object which responds to inquiry for concept name (associated with Concept_Number), and the possibility of linking back to the sections of the student assignment which are good, and those needing improvement.

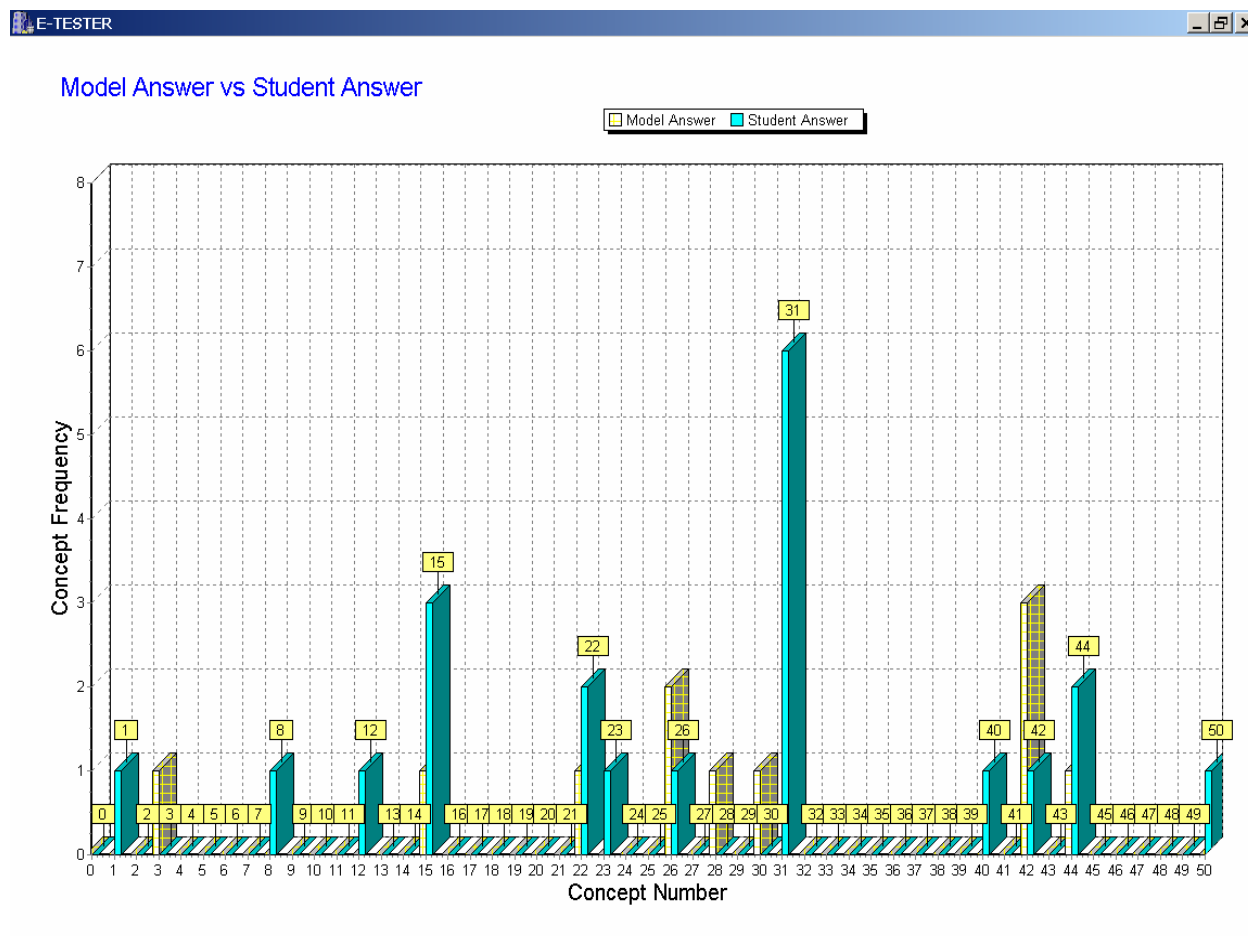


Figure 4 – Concept frequencies: student answer and course content

It is proposed to further develop MarkIT so that these graphs, in computer readable format, will form part of the feedback to the student and the teacher. The teacher will then be able to interactively explain to the student the strengths and weaknesses of the student's answer. If a bar in the graph is double clicked, the thesaurus text for the category represented by the bar will be displayed. The student can then see the amount of discussion that should have been devoted to the topic, and also get a good feel, from the many words in that thesaurus category, how to express that content. A percentage of the discussion above or below the expected amount of discussion will also be displayed.

Summary

MarkIT has been developed to provide automated grading of essay-type documents. Along with its peers in the AEG domain MarkIT performs as well as human graders under certain given conditions. Unlike many of its competitors, MarkIT is now endowed with the added feature of providing meaningful, relevant, and detailed feedback to assist learners improve their performance.

References

Attali, Y. & Burstein, J. (2004). Automated essay scoring with E-rater V.2.0. Paper presented at the *Conference of the International Association for Educational Assessment (IAEA)*, June 13-18, 2004, Philadelphia, USA. Retrieved from <http://www.ets.org/research/dload/IAEA.pdf>

- Callan, J. P., Croft, W. B. & Broglio, J. (1995). TREC and TIPSTER experiments with INQUERY. *Information Processing and Management*, 327-343.
- College Board (2002). The new SAT: Implemented for the class of '06. PowerPoint slides posted on http://www.collegeboard.com/prod_downloads/about/newsat/newsat_presentation.ppt
- Dreher, H., Scerbakov, N., & Helic, D. (2004). Thematic driven learning. *Proceedings of E-Learn 2004 Conference*, Washington DC, USA, November 1-5. Retrieved from <http://www.aace.org/conf/ELearn/>
- Foltz, P., Laham, D. & Landauer, T. (1999). Automated essay scoring: Applications to educational technology. Retrieved from <http://www-psych.nmsu.edu/~pfoltz/reprints/Edmedia99.html>
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.
- Larkey, L. S. (1998). Automatic essay grading using text categorization techniques, *Proceedings of the Twenty First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 90-95.
- Maron, M. E. (1961). Automatic indexing: An experimental inquiry. *Journal of the Association for Computing Machinery*, 8, 404-417.
- Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, January, 238-243.
- Page, E. B. (1994). Computer grading of student prose, using modern concepts and software, *Journal of Experimental Education*, 62, 127-142.
- Shermis, M. & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective 2003*. New Jersey, USA: Lawrence Erlbaum Associates.
- Shermis, M., Mzumara, H., Olson, J. & Harrington, S. (2001). On-line grading of student essays: PEG goes on the World Wide Web. *Assessment and Evaluation in Higher Education*, 26, 3. Retrieved from <http://taylorandfrancis.metapress.com/media/804PYKUXVJC2076KLQFW/Contributions/H/E/8/4/HE84J5VPEDVRVL3T.pdf>
- Valenti, S., Neri F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2, 319 – 330. Retrieved from <http://jite.org/documents/Vol2/v2p319-330-30.pdf>
- van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). London: Butterworths.
- Williams, R. (2001). Automated essay grading: An evaluation of four conceptual models. In M. Kulski & A. Herrmann (Eds.), *New horizons in university teaching and learning: Responding to change*. Perth, Australia: Curtin University of Technology.
- Williams, R. & Dreher, H. (2004). Automatically grading essays with Markit©. *Journal of Issues in Informing Science and Information Technology*, 1, pp693-700

Biographies



Robert Williams has over 25 year's experience in the Information Systems industry, as a practitioner, researcher and lecturer. He currently is a lecturer in the School of Information Systems at Curtin University of Technology in Perth, Western Australia. He has extensive experience in systems analysis and design, and programming, on a variety of mainframe, mini and personal computers, and a variety of operating systems and programming languages. Applications he has worked with include mathematical, statistical, bridge and road engineering, financial, corporate resource allocation, business simulation and educational systems. He has published a number of articles on system users' personalities and satisfaction, decision support systems, and automated essay grading

systems. In 2001 he led a team of researchers in the School of Information Systems at Curtin University of Technology which conducted what is believed to be the first trial in Australia of an Automated Essay Grading system. Robert holds a Bachelor of Arts degree with double majors in Mathematics and Economics from the University of Western Australia, a Graduate Diploma in Computing from the Western Australian Institute of Technology, and a Master of Information Systems degree from Curtin University of Technology.



Heinz Dreher is senior lecturer and research fellow in the School of Information Systems at Curtin University of Technology. He has published in the educational technology and information systems domain through conferences, journals, invited talks and seminars; is currently the holder of Australian National Competitive Grant funding for a 4 year e-Learning project; is participating in a Digital Library project with TU Graz / Austria; is collaborating on Automated Essay Grading technology development, trial usage and evaluation; has received numerous industry grants for investigating hypertext based systems in training and business scenarios; and is an experienced and accomplished teacher, receiving awards for his work in cross-cultural awareness and course design. In 2004 he was appointed Adjunct Professor for Computer Science at TU Graz, and continues to collaborate in teaching & learning, and research projects with local and overseas partners.

Paper No. 5

Williams, R. (2005). *MarkIT in Action - A Report on the Automatic Grading of DET Year 10 Essays*, Curtin University of Technology, Perth, Australia.

This paper is unable to be reproduced here due to proprietary reasons.

Williams, R. (2006). The Power of Normalised Word Vectors for Automatically Grading Essays. *Journal of Issues in Informing Science and Information Technology*, Vol. 3, pp. 721-729.

The Power of Normalised Word Vectors for Automatically Grading Essays

Robert Williams

***School of Information Systems, Curtin University of Technology
Perth, Australia***

Bob.Williams@cbs.curtin.edu.au

Abstract

Latent Semantic Analysis, when used for automated essay grading, makes use of document word count vectors for scoring the essays against domain knowledge. Words in the domain knowledge documents and essays are counted, and Singular Value Decomposition is undertaken to reduce the dimensions of the semantic space. Near neighbour vector cosines and other variables are used to calculate an essay score. This paper discusses a technique for computing word count vectors where the words are first normalised using thesaurus concept index numbers. This approach leads to a vector space of 812 dimensions, does not require Singular Value Decomposition, and leads to a reduced computational load. The cosine between the vectors for the student essay and a model answer proves to be a very powerful independent variable when used in regression analysis to score essays. An example of its use in practice is discussed.

Keywords: Automated Essay Grading, Latent Semantic Analysis, Singular Value Decomposition, Normalised Word Vectors, Electronic Thesaurus, Multiple Regression Analysis.

Introduction

Automated Essay Grading (AEG) systems are now appearing in the educational marketplace, and are increasingly being accepted as a way of efficiently grading large numbers of essays (Shermis & Burstein, 2003). There are many theoretical constructs underpinning the various AEG systems (Williams, 2001; Valenti, Neri & Cucchiarelli, 2003). One of the major systems, the Intelligent Essay Assessor (Pearson Knowledge Technologies, 2005), makes use of a mathematical technique known as Latent Semantic Analysis (LSA) (Landauer, Foltz & Laham, 1998). This system is interesting because of the way it derives the knowledge contained in an essay from the words comprising the essay. The MarkIT system (Williams & Dreher, 2005), being developed by the author and colleagues, uses an alternative way of deriving content from an essay, but still based on the words making up the essay. This paper discusses these two alternative word-based content representations, presents new material on the grading algorithm for MarkIT, and compares the performances of the two systems.

Material published as part of this publication, either on-line or in print, is copyrighted by the Informing Science Institute. Permission to make digital or paper copy of part or all of these works for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage AND that copies 1) bear this notice in full and 2) give the full citation on the first page. It is permissible to abstract these works so long as credit is given. To copy in all other cases or to republish or to post on a server or to redistribute to lists requires specific permission and payment of a fee. Contact Publisher@InformingScience.org to request redistribution permission.

In this paper we do not have space to give a detailed coverage of the issues associated with AEG systems. For a comprehensive coverage of AEG systems, their algorithms, and performance details, see Hearst (2000), Williams (2001), and Valenti, Neri and Cucchiarelli (2003).

Latent Semantic Analysis

LSA is a mathematical technique based on vector algebra. It is used to derive a representation of the content of a collection of text documents in a particular domain of knowledge. This content representation is generally termed the semantic space. This space is built from text segments that may consist of the complete documents, or subsets of the documents, such as paragraphs or sentences. Each word in the segment is represented as a row in a matrix, and each segment is represented as a column in the same matrix. The counts of the number of times the words appear in the segments are entered in the corresponding elements in the matrix.

The following example, taken from Landauer, Foltz, and Laham (1998) and used with permission from the authors and Lawrence Erlbaum Associates, the publishers, illustrates the technique. The titles of five documents relating to human computer interaction and four relating to mathematical graph theory are shown below.

- c1: *Human machine interface* for ABC computer applications
- c2: A survey of *user* opinion of *computer system response time*
- c3: The *EPS user interface* management system
- c4: *System* and *human system* engineering testing of *EPS*
- c5: Relation of *user* perceived *response time* to error measurement
- m1: The generation of random, binary, ordered *trees*
- m2: The intersection *graph* of paths in *trees*
- m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
- m4: *Graph minors*: A survey

The matrix below shows the word count for the selected words occurring in at least two of the titles. These words are shown in *italics* in the document titles.

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

A vector algebra technique, known as Singular value Decomposition (SVD) is then applied to this matrix. SVD breaks the matrix into 3 component matrices that can be matrix multiplied to produce the original matrix. However the dimensions of these 3 matrices are reduced before the remultiplication. The remultiplied matrix is now approximately equivalent to the original matrix in terms of its element values, but now contains values for elements that were previously zero. In other words, the reconstituted matrix now has relationships for words and segments that were not explicitly displayed in the original matrix, but have been induced by the SVD process from the hidden or latent relationships amongst the words and segments. The reconstructed approximation to the original matrix, based upon the first two columns in the three component matrices (not shown), is

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

What was originally a sparsely populated matrix of relationships amongst words and segments is now a rich array of associations. This is now the semantic space for this collection of document titles.

“This text segment is best described as having so much of abstract concept one and so much of abstract concept two, and this word has so much of concept one and so much of concept two, and combining those two pieces of information (by vector arithmetic), my best guess is that word X actually appeared 0.6 times in context Y.” (Landauer, et al., 1998, p 264)

Essays on a particular topic are graded as follows. The appropriate semantic space is built – this can be done by processing electronic texts on the topic, or from a collection of several hundred human graded essays on the topic. The essay to be graded is then processed using the SVD technique to build a document vector in this space. An essay score is then computed from near neighbour human scored essay vectors in this space, and other variables.

The IEA is a commercial implementation of the LSA approach to AEG. Landauer indicates that this system builds the semantic space as follows:

“IEA/LSA always starts from a reduced dimensional space based on a large relevant corpus to which it adds text special to the topic and the student essays” (personal email communication, 16 November, 2005).

Evaluation of LSA and Essay Grading

Nichols has evaluated the IEA. He concludes

“All four of the measures of the relationship between essay scores and expert scores (percent agreement, Spearman rank-order correlation, kappa statistic and Pearson correlation) indicated a stronger relationship between the IEA and experts than between readers and experts. In addition, the results of examining the scoring processes used by the IEA showed that the IEA used processes similar to a human scorer. Furthermore, the IEA scoring processes were more similar to processes used by proficient human scorers than to processes used by non-proficient or intermediate human scorers.” (Nichols, 2005, p 21).

Vector Representation of Documents using a Thesaurus to Normalise Document Words

The MarkIT AEG system is a software system that automatically grades essays against an ideal content answer at the same level of accuracy as human graders (MarkIT, 2005; Williams & Dre-

her, 2005). This section explains how vector algebra techniques are used to represent similarities in content between documents in MarkIT. In order to build this vector representation, a thesaurus is used to “normalise” words in the documents by reducing all words to a thesaurus root word appropriate to the concept the word belongs to. Counts of these concepts are then used for the vector representation. Consider the following start of sentence fragments from successive sentences in 3 separate documents:

<u>Document Number</u>	<u>Document Text</u>
(1)	The little boy... A small male...
(2)	A minor boy... A funny girl...
(3)	The large boy... Some minor day...

Suppose a thesaurus exists with the following root concept numbers and words:

<u>Concept Number</u>	<u>Words</u>
1.	the, a
2.	little, small, minor
3.	boy, male
4.	large
5.	funny
6.	girl
7.	some
8.	day

Three dimensional vector representations of the above document fragments on the first 3 concept numbers (1-3) can be constructed by counting the number of times a word in that concept number appears in the document fragments. These vectors are:

<u>Document Number</u>	<u>Vector on first 3 concepts</u>	<u>Explanation</u>
(1)	[2, 2, 2]	[The, a; little, small; boy, male]
(2)	[2, 0, 1]	[A, a; ; boy]
(3)	[1, 1, 1]	[The; minor; boy]

Figure 1 shows these 3 dimensional vectors pictorially.

Computing the Variable CosTheta

If we assume that document 1 is the model answer, then we can see how close semantically documents 2 and 3 are to the model answer by looking at the closeness of their corresponding vectors. The angle between the vectors varies according to how “close” the vectors are. A small angle indicates that the documents contain similar content, a large angle indicates that they do not have much common content. Angle Theta1 is the angle between the model answer vector and the vector for document 2, and angle Theta2 is the angle between the model answer vector and the vector for document 3.

The cosines of Theta1 and Theta2 can be used as measures of this closeness. If documents 2 and 3 were identical to the model answer, their vectors would be identical to the model answer vector, and would be collinear with it, and have a cosine of 1. If on the other hand, they were completely different, and therefore orthogonal to the model answer vector, their cosines would be 0.

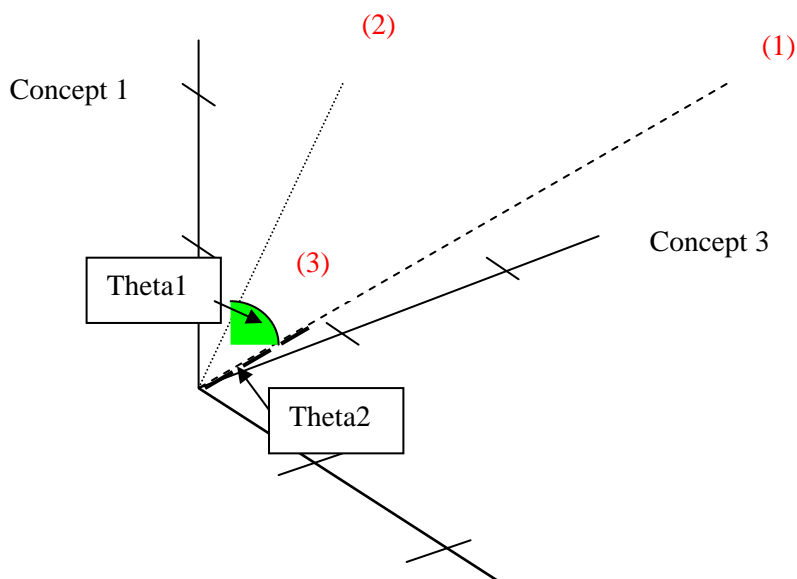


Figure 1. Vector representation (dashed lines) of documents

Generally in practice, a document's cosine is between these upper and lower limits.

The variable CosTheta used in the scoring algorithm is this cosine computed for the document being scored.

In general, these ideas are extended to the 812 concepts in the Macquarie Thesaurus from Macquarie Library Pty Ltd (Macquarie Library, 2005), and all words in the documents. This means that the vectors are constructed in 812 dimensions, and the vector theory carries over to these dimensions in exactly the same way – it is of course hard to visualise the vectors in this hyperspace. (The system developers approached a number of thesaurus publishers with a view to obtaining a research licence to use an electronic thesaurus, and Macquarie Library Pty Ltd was the only company willing to grant one; hence its usage).

Computing the Variable VarRatio

We now discuss another powerful essay grade predictor, VarRatio , which is based on these concept vectors. The number of concepts that are present in the model answer (document 1) above is 3. This can be determined from the number of non-zero counts in the numerical vector representation.

The number of concepts that are present in document 2 above is 2 – the second vector index is 0. To compute the VarRatio for this document 2 we divide the non-zero concept count for document 2 by the non-zero concept count in the model answer i.e. $\text{VarRatio} = 2/3 = 0.67$. The corresponding VarRatio for document 3 is $3/3 = 1.00$.

This simple variable provides a remarkably strong predictor of essay scores, and is generally present as one of the components in the scoring algorithm.

Scoring Student Essays by Matching a Model Answer against Student Answers

MarkIT makes use of a multiple regression equation to assign a grade to a student essay. The regression equation is developed from about 100 human graded training essays and an ideal or model answer. The document vectors described above are constructed. Values are then computed for many variables from the relationships between the content and vectors of the model answer and the training essays. Once the training has been performed, and the grading algorithm built, each unmarked essay is processed to obtain the values for the independent variables, and the regression equation is then applied. Generally CosTheta and VarRatio are significant predictors in the scoring equation. An example taken from a trial of the system is now discussed.

In the trial, Year 10 high school students hand wrote essays on paper on the topic of “The School Leaving Age”. Three trained human graders then graded these essays against a marking rubric. The essays, 390 in total, were then transcribed to Microsoft Word document format. The essay with the highest average human score was selected as the model answer. It had a score of 48.5 out of a possible 54, or 90%. In one test of the system, 100 essays were used to build the scoring algorithm. The scoring algorithm was built using the first 100 essays in the trial when ordered in ascending order of the identifier. Table 1 shows the results of the multiple regression procedure built upon the output of the MarkIT system for these 100 essays. The multiple R is 0.89 and the prediction equation is

$$\text{Student Grade} = -22.35 + 11.00 \cdot \text{CosTheta} + 15.70 \cdot \text{VarRatio} + 7.64 \cdot \text{Characters Per Word} + 0.20 \cdot \text{Number of NP Adjectives}$$

Table 1. Multiple Regression Analysis for First 100 Essays

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.89
R Square	0.79
Adjusted R Square	0.78
Standard Error	4.16
Observations	100.00

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	4.00	6079.76	1519.94	87.71	0.00
Residual	95.00	1646.21	17.33		
Total	99.00	7725.97			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	-22.35	6.67	-3.35	0.00
CosTheta	11.00	3.74	2.94	0.00
VarRatio	15.70	2.86	5.49	0.00
Characters Per Word	7.64	1.74	4.40	0.00
Number of NP Adjectives	0.20	0.08	2.41	0.02

- CosTheta is computed as per the explanation above.
- VarRatio is computed as per the explanation above.
- Characters Per Word is the average number of characters in the words in the essay
- Number of NP Adjectives is the number of Adjectives in Noun Phrases in the essay

Notice that only 4 independent variables are needed for the predictor equation in this example.

Once this scoring algorithm was coded into the scoring program, the remaining 290 essays were graded by it. Figure 2 shows the results.

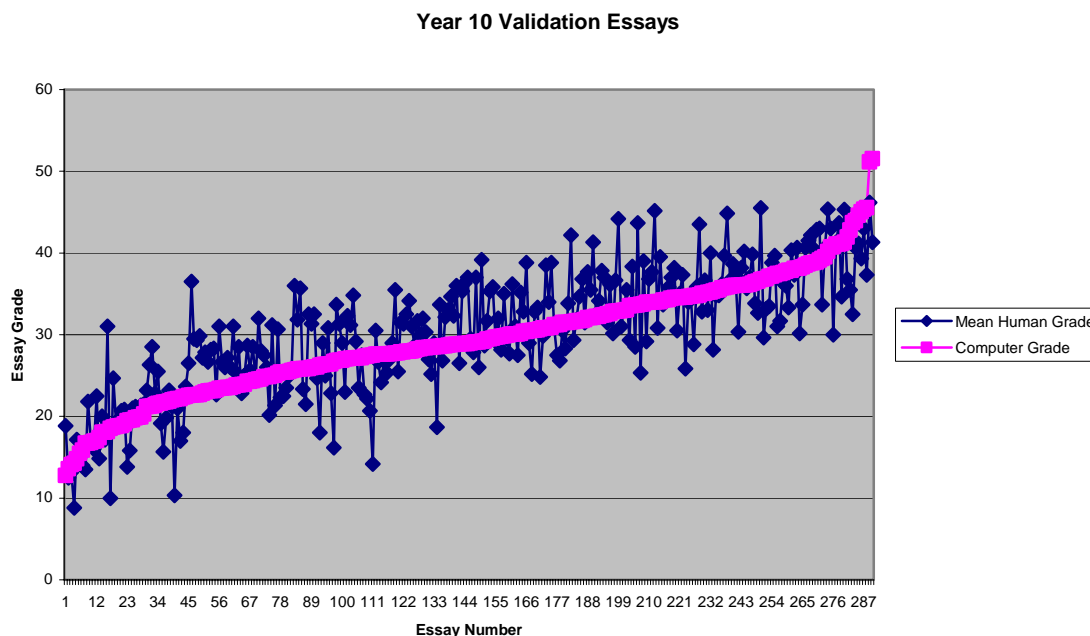


Figure 2. Results of Computer Scoring of Last 290 Essays

The mean score for the human average grade for these 290 essays was 30.34, while the mean grade given by the computer was 29.45, a difference of 0.89. The correlation between the human and computer grades was 0.79. The mean absolute difference between the two was 3.90, representing an average error rate of 7.23% when scored out of 54 (the maximum possible human score).

The correlations between the three humans amongst themselves were 0.81, 0.78 and 0.81.

The benefits of averaging the scores from the human graders are shown by the fact that the correlation between the computer and the mean score of the three humans is higher, at 0.79, than the individual correlations at 0.67, 0.75 and 0.75.

Conclusion

LSA makes use of SVD to reduce the large number of dimensions generated when each word in a document is counted as a separate dimension. Typically the dimensions are reduced to about 300 (Landauer, 2005). The processing involved for the SVD takes a few hours on a common small Linux cluster (Landauer, personal email communication, 16 November 2005).

While the number of dimensions resulting from normalising words against thesaurus index numbers is 812, much less processing is involved – typically the training session to build the scoring algorithm for a prompt using 100 essays takes 5 minutes on a Pentium 3.4GHz machine under Windows XP. Similar accuracy of the resultant scores, when compared to human scores, is maintained. For example, IEA achieved a correlation of 0.81 with human scores for GMAT essays (Landauer, Laham & Foltz, 2003), compared to the 0.79 achieved by MarkIT for the Year 10 High School essays reported above.

The power of the resultant document vectors to represent the essay content is also impressive, as only the cosine of the model and student essay vectors, and three other predictors, are needed for scoring the student essay, in the example discussed. This low number of predictors appears to be unique to MarkIT. Other documented systems appear to require substantially more (Shermis & Burstein, 2003).

Acknowledgement

The author wishes to express his appreciation to Professor Tom Landauer of the University of Colorado, Boulder, for his valuable comments and suggestions for improvements during the writing this paper.

References

- Hearst, M. (2000). The debate on automated essay grading. *IEEE Intelligent Systems*, 15 (5), 22-37, IEEE CS Press.
- Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.
- Landauer, T., Laham, D. & Foltz, P. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor™. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp 87–112). New Jersey, USA: Lawrence Erlbaum Associates.
- Macquarie Library. (2005). Retrieved November 29, 2005 from <http://www.macquariedictionary.com.au/>
- MarkIT (2005). Retrieved November 28, 2005 from <http://www.essaygrading.com/>
- Nichols, P. (2005). Evidence for the interpretation and use of scores from an automated essay scorer. *PEM Research Report 05-02*, Retrieved November 28, 2005 from http://www.pearsonedmeasurement.com/downloads/research/RR_05_02.pdf
- Pearson Knowledge Technologies (2005). Retrieved November 28, 2005 from <http://www.knowledge-technologies.com/>
- Shermis, M. & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective*. New Jersey, USA: Lawrence Erlbaum Associates.
- Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading, *Journal of Information Technology Education*, 2, 319 – 330. Retrieved January 30, 2006 from <http://jite.org/documents/Vol2/v2p319-330-30.pdf>
- Williams, R. (2001) Automated essay grading: an evaluation of four conceptual models. In M. Kulski & A. Herrmann (Eds.), *New horizons in university teaching and learning: Responding to change*. Perth, Australia: Curtin University of Technology.
- Williams, R. & Dreher, H. (2005). Formative assessment visual feedback in computer graded essays. *Journal of Issues in Informing Science and Information Technology*, 2, 23-32. Retrieved from <http://proceedings.informingscience.org/InSITE2005/I03f95Will.pdf>

Biography



Robert Williams has over 25 year's experience in the Information Systems industry, as a practitioner, researcher and lecturer. He currently is a lecturer in the School of Information Systems at Curtin University of Technology in Perth, Western Australia. He has extensive experience in systems analysis and design, and programming, on a variety of mainframe, mini and personal computers, and a variety of operating systems and programming languages. Applications he has worked with include mathematical, statistical, bridge and road engineering, financial, corporate resource allocation, business simulation and educational systems. He has published a number of articles on system users' personalities and satisfaction, decision support systems, and automated essay grading systems. In 2001 he led a team of researchers in the School of Information Systems at Curtin University of Technology which conducted what is believed to be the first trial in Australia of an Automated Essay Grading system. Robert holds a Bachelor of Arts degree with double majors in Mathematics and Economics from the University of Western Australia, a Graduate Diploma in Computing from the Western Australian Institute of Technology, and a Master of Information Systems degree from Curtin University of Technology.

Williams, R. (2007). A Computational Effective Document Semantic Representation. *Proceedings of IEEE-Digital Ecosystems and Technologies 2007 Conference*, Cairns, Australia, 21-23 February, pp. 410-415.

This material is posted here with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of the Curtin University of Technology's products or services. Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by writing to pubs-permissions@ieee.org. By choosing to view this material, you agree to all provisions of the copyright laws protecting it.

A Computational Effective Document Semantic Representation

© 2007 IEEE. Reprinted, with permission, from the Proceedings of the 2007 IEEE International Conference on Digital Ecosystems and Technologies, A Computational Effective Document Semantic Representation, Robert Williams.

Robert Williams

School of Information Systems, Curtin University of Technology, Perth, Western Australia

e-mail: bob.williams@cbs.curtin.edu.au

Abstract— A technique based on Noun Phrase and Verb Clause slot structures is described for representing the semantics of the sentences making up a text document. Thesaurus head word index numbers are placed in the appropriate document sentence clause slots to represent the meta level meaning of the sentences. Many different expressions of the same document content can thus be represented by one semantic representation. An implementation of such a technique is described, and sample output is presented. The document summarisation thus produced is suitable for manipulation by computers for a variety of document processing tasks. The technique has primarily been developed for an Automated Essay Grading system, where a robust context free representation of documents is required.

Index Terms—document semantic representation, thesaurus, meta level meaning, document summarisation, automated essay grading.

I. INTRODUCTION

Representing the semantics of a text document for computational uses is problematic. How can we formally code the meanings of words, phrases, sentences, paragraphs and so on, so that useful computational work can be done robustly, effectively and efficiently? The computational work may involve text understanding, question answering, or essay grading to name a few possible applications.

In this article we discuss one such technique that is being used in a system being developed for Automated Essay Grading (AEG).

The technique allows a formal representation of free unseen text to be quickly and robustly built for further analysis by the AEG system.

II. CONTEMPORARY SEMANTIC REPRESENTATION

Poesio [6] discusses some current techniques for representing the meaning of sentences, including First-Order Logic and Semantic Networks. First-Order Logic uses mathematical expressions representing set membership relations for objects in the sentence belonging to sets to represent the meaning of a sentence. Computationally, this is very difficult to use for unlimited unseen text, as generally domain specific information needs to be hand coded. Semantic Networks use classifications of objects into a network of relationships, and arc traversal of the nodes can be used to imply relationships amongst the nodes. Again, substantial

domain specific knowledge needs to be hand coded prior to their use.

There is a need for a semantic representation that does not need substantial hand coding of knowledge structures prior to use, and that can deal with unlimited unseen text.

III. PRACTICAL LIMITATIONS OF CONTEXT FREE PHRASE STRUCTURE PARSERS FOR PRELIMINARY PROCESSING

Many Natural Language Processing (NLP) systems use some kind of a parser to initially extract the syntax of sentences in a document as an initial step prior to further processing. Semantic analysis then follows. The use of Context Free Phrase Structure Grammar (CFPSG) parsers is commonly suggested in the literature. They require an extensive set of grammar rules which define legitimate syntax structures. However it is virtually impossible to build a set of grammar rules for free unseen text in practice, because thousands of grammar rules are typically required, and over-generation of possible parse trees results. Increases in parsing time become exponential as the parse trees proliferate. So context free CFPSG parsing cannot be used in all but simple toy domains.

IV. CHUNKING AS AN ALTERNATIVE TO FULL PARSING

Useful preliminary linguistic computation can be done with less structured parsing. Phrase chunking can be an effective alternative to full parsing at the initial processing stage. Chunking has the advantage that it does not require an extensive set of grammar rules – a few simple rules suffice. Specifically chunking breaks a sentence into syntactically structured components representing noun clauses or phrases and verb clauses or phrases. Often, these structures are sufficient as a preliminary to further processing.

The technique outlined in this paper uses chunking to extract noun phrase and verb clause structures for further processing.

V. STRUCTURES TO HOLD CHUNK SEMANTIC DETAILS

The technique described in this paper makes use of chunking to get the structure of sentences in terms of subject and predicate, as represented by Noun Phrases (NP) and Verb Phrases (VP). Generally the NP nominates the subject of discussion, and the VP the actions being performed on or by the subject. However VPs are notoriously complex to deal with in comparison to NPs, because they typically can

have many clusters of a Verb Clause (VC) and a NP together. It is far easier to identify VCs instead of the complex VPs. The basis of the technique used is to represent the meaning of the words making up the NPs and VCs in a sequence of structured slots containing a numerical value representing the thesaurus index number for the root meaning of the word in the slot. A numerical summary of the meaning of the sentences in the document being considered is thus built up.

The exact structure of the NP and VC slots is discussed further below, but to illustrate the concept and to give a practical example, consider the following. A typical sentence would comprise alternating NPs and VCs as follows. A typical first NP slot word and numerical contents would be

DET ADJ ADJ N
The small black dog
100 143 97 678

A typical first VC slot word and numerical contents would be

V ADV ADV
walked slowly down
34 987 67

A typical concluding NP slot word and numerical contents would be

DET N
the street
100 234

where the numbers are the thesaurus index numbers for the corresponding words. The numbers here are fictitious, for illustration purposes only. A sentence generally consists of groups of alternating NPs and VCs, not necessarily in that order, so a sentence summary would be represented by a group of NP slots and VC slots containing numerical thesaurus indices. A document summary would then consist of a collection of these groups. Note that a sentence does not have to start with a NP, but can start equally well with a VP.

A. Proposed NP Structure

Martha Kolln [2] on page 433 states a rule for defining an NP under transformational grammar as follows

(1) NP = (DET) + (ADJ) + N +(PREP PHR) + (S)

and on page 429 a Prep Phr as follows

PREP PHR = PREP + NP

When considering the slots to be provided for a NP, (1) above can now be rewritten as

(2) NP = DET ADJ N PREP NP S

The basic component of an NP appears to be

(3) NP = DET ADJ N and some appended structures. It has been found in practice that

(4) NP = DET ADJ ADJ ADJ N

to be a better structure. If we take this as a basic core structure in a NP, the complete NP structure can be built in terms of this core structure by linking multiple occurrences of this core structure by PREPs. It has been found in practice that we should also allow linking by CONJs. So finally we conclude that the basic component should be

(5) NP = CONJ PREP : DET ADJ ADJ ADJ N

where the 2 slots before the colon are the linking slots, and those following the content slots. Practice indicates that we should allow about 40 occurrences of this basic component as the NP slot template should handle many practical NPs encountered in general English text. So a 40x7 array with the following structure will be needed in the program. Fig. 1 shows the first 10 rows of this array.

CONJ	PREP	DET	ADJ	ADJ	ADJ	N

Fig. 1. Noun Phrase Semantic Structure

The first core component in the sentence generally will have the CONJ and PREP slots set to blank (in fact the number 0). Any empty slots will likewise be set to 0.

B. Proposed VC Structure

Martha Kolln [2] on page 428 states a rule for defining a VP under transformational grammar as follows

(6) VP = AUX + V + (COMP) + (ADV)

COMP is explained as an NP or ADJ, so by removing this from the VP we end up with a VC as follows

(7) VC = AUX + V + ADV

It has been found in practice that if we modify this VC definition by the addition of extra AUXs and ADVs we obtain a more useful structure as

(8) VC = AUX AUX ADV ADV V AUX AUX ADV ADV

Vcs can often be introduced with CONJs, and it has been found in practice that we should also allow PREPs in a VC, so a complete VC definition would be

(9) VC = CONJ PREP AUX AUX ADV ADV V AUX
AUX ADV ADV

If we allow for 40 occurrences of this basic VC component to handle VCs encountered in practice, we will need the following 40x11 array structure in the program. Fig. 2 shows the first 10 rows of this array.

C	P	A	A	AD	AD	V	A	A	A	A
O	R	U	U	V	V		U	U	D	D
NJ	E	X	X				X	X	V	V

Fig. 2. Verb Clause Semantic Structure

If a sentence happens to start with a VC, then the CONJ slot will be set to blank (in fact the number 0). Any empty slots will likewise be set to 0.

VI. ALGORITHMS FOR BUILDING A CHUNKED SEMANTIC REPRESENTATION

It is all well and good to postulate a theoretical model of semantic representation, but can it be implemented in a practical way? The answer is yes, and details of the author's implementation of the concepts are discussed below.

The following algorithm describes the process.

```

For each sentence in the document
  Tag each word with POS
  Convert POS tags to standard format
  Stem each word
  For each word and/or stem
    Extract thesaurus indices and POS (many)
    Determine within context thesaurus index and POS (one)
    Store thesaurus index, POS, Word
  Chunk sentence
  Store chunks in NC and VC slots
End

```

The following sentence will be used as an example in the explanation that follows. It has been chosen for its relative complexity, to show that the system can handle more than trivial sentences.

“For example if people working on a group project did work their own way and on their own schedule it would be extremely difficult to coordinate their work and assure the quality and timeliness of the end product.”

(Source: [1])

A. Tag Each Word with POS

As with many NLP systems, we start with Part of Speech (POS) tagging of the words in the sentence, one sentence at a time. This allows the system to have a preliminary understanding of the words in the sentence it will be dealing with. The Qtag tagger from Mason [3] is currently used. It produces

```

<w pos="IN">For</w>
<w pos="RB22">example</w>
<w pos="CS">if</w>
<w pos="NN">people</w>
<w pos="VBG">working</w>
<w pos="IN">on</w>
<w pos="DT">a</w>
<w pos="NN">group</w>
<w pos="NN">project</w>
<w pos="DOD">did</w>
<w pos="NN">work</w>
<w pos="PP$">their</w>
<w pos="DT">own</w>
<w pos="NN">way</w>
<w pos="CC">and</w>
<w pos="IN">on</w>
<w pos="PP$">their</w>
<w pos="DT">own</w>
<w pos="NN">schedule</w>
<w pos="PP">it</w>
<w pos="MD">would</w>
<w pos="BE">be</w>
<w pos="RB">extremely</w>
<w pos="JJ">difficult</w>
<w pos="IN">to</w>
<w pos="VB">coordinate</w>
<w pos="PP$">their</w>
<w pos="VB">work</w>
<w pos="CC">and</w>
<w pos="VB">assure</w>
<w pos="DT">the</w>
<w pos="NN">quality</w>
<w pos="CC">and</w>
<w pos="NN">timeliness</w>
<w pos="IN">of</w>
<w pos="DT">the</w>
<w pos="NN">end</w>
<w pos="NN">product</w>

```

B. Convert POS Tags to a Standard Format

To reduce the number of tags the system has to deal with, the numerous tags produced by Qtag are reduced to a standard set, which eases the computational load in later processing. The system changes the above tag information to the following.

P For
 N example
 CONJ if
 N people
 V working
 P on
 DET a
 N group
 N project
 AUX did
 N work
 N their
 DET own
 N way
 CONJ and
 P on
 N their
 DET own
 N schedule
 N it
 AUX would
 AUX be
 ADV extremely
 ADJ difficult
 P to
 V coordinate
 N their
 V work
 CONJ and
 V assure
 DET the
 N quality
 CONJ and
 N timeliness
 P of
 DET the
 N end
 ADJ product

C. Stem each Word

The words produced above will be input to a database containing an electronic version of a thesaurus to attempt to find a head word index number. Additional words, particularly conjunctions and prepositions have been added to this document to rectify their omission from a standard thesaurus. These are represented by index numbers over 1000. If the word cannot be found in the thesaurus, the word's stem is input in attempt to find the word's base form. Many words, such as 'working' do not appear in the thesaurus, but its stem 'work' does. In this case we use the thesaurus index number for 'work' instead of 'working', without losing substantial meaning of the word.

The stemming program used is an implementation of the Porter stemming algorithm documented in [5]. It produces the following output.

for
 exampl
 if
 peopl

work
 on
 a
 group
 project
 did
 work
 their
 own
 wai
 and
 on
 their
 own
 schedul
 it
 would
 be
 extrem
 difficult
 to
 coordin
 their
 work
 and
 assur
 the
 qualiti
 and
 timeli
 of
 the
 end
 product

D. For each Word and/or Stem

a) Extract Thesaurus Indices and POS (many)

We now extract the POS and head word index numbers from the thesaurus. Only the POS that matches the POS for the input word is output. This process produces the following output. Notice that many words have multiple entries. Eg 'working'. An index number of 8888 indicates that an entry could not be found for the word in the thesaurus.

3027 P For
 22 N example
 4012 CONJ if
 997 N people
 677 V working
 680 V working
 686 V working
 3034 P on
 2000 DET a
 712 N group
 8888 N project
 5008 AUX did
 154 N work
 170 N work
 415 N work

593 N work
 625 N work
 686 N work
 2008 N their
 8888 DET own
 26 N way
 180 N way
 627 N way
 4003 CONJ and
 3034 P on
 2008 N their
 8888 DET own
 86 N schedule
 8888 N it
 5032 AUX would
 5002 AUX be
 8888 ADV extremely
 868 ADJ difficult
 3046 P to
 60 V coordinate
 2008 N their
 677 V work
 680 V work
 686 V work
 4003 CONJ and
 858 V assure
 2007 DET the
 5 N quality
 157 N quality
 812 N quality
 875 N quality
 4003 CONJ and
 8888 N timeliness
 3032 P of
 2007 DET the
 620 N end
 8888 ADJ product

E. For each Word and/or Stem

b) Determine within Context Thesaurus Index and POS (one). Store Thesaurus Index, POS, Word

As can be seen, many words have multiple entries in the output above. This process now selects the most appropriate entry by using a 'within context' algorithm. The entry chosen is the one which makes the most sense in the context of the other words in the sentence. This is done by using broader groups of word categories that are indicated in the related words of the thesaurus classification.

These processes produce the following output. Notice that 'working' now has only one entry.

3027 P For
 22 N example
 4012 CONJ if
 997 N people
 677 V working
 3034 P on
 2000 DET a

712 N group
 8888 N project
 5008 AUX did
 154 N work
 2008 N their
 8888 DET own
 26 N way
 4003 CONJ and
 3034 P on
 2008 N their
 8888 DET own
 86 N schedule
 8888 N it
 5032 AUX would
 5002 AUX be
 8888 ADV extremely
 868 ADJ difficult
 3046 P to
 60 V coordinate
 2008 N their
 677 V work
 4003 CONJ and
 858 V assure
 2007 DET the
 5 N quality
 4003 CONJ and
 8888 N timeliness
 3032 P of
 2007 DET the
 620 N end
 8888 ADJ product

F. Chunk Sentence

The above output is now input into the chunking process. This process uses generic sequences of POS to determine the start of NPs and VCs, and then fills the slots for the clauses with the composing words and index numbers. The object-oriented context free Phrase Structure Grammar parser written in C++ described by Perelman-Hall [4] has been substantially adapted to implement the concepts described previously. This process produces the following output. Slots containing blanks and zeroes have been eliminated because of space limitations.

NOUN PHRASE

FOR EXAMPLE
 IF PEOPLE

0 3027 0 0 0 0 22 0
 4012 0 0 0 0 0 997 0

VERB PHRASE

WORKING

0 0 0 0 0 0 677 0 0 0 0 0

NOUN PHRASE

ON A GROUP

PROJECT

0 3034 2000 0 0 0 712 0
0 0 0 0 0 0 8888 0

VERB PHRASE

DID WORK

0 0 5008 0 0 0 677 0 0 0 0 0

NOUN PHRASE

THEIR
OWN WAY
AND ON THEIR
OWN SCHEDULE
IT

0 0 0 0 0 0 2008 0
0 0 8888 0 0 0 26 0
4003 3034 0 0 0 0 2008 0
0 0 8888 0 0 0 86 0
0 0 0 0 0 0 8888 0

VERB PHRASE

WOULD BE EXTREMELY

0 5032 5002 8888 0 0 0 0 0 0 0

NOUN PHRASE

TO DIFFICULT

0 3046 0 868 0 0 0 0

VERB PHRASE

COORDINATE

0 0 0 0 0 0 60 0 0 0 0 0

NOUN PHRASE

THEIR

0 0 0 0 0 0 2008 0

VERB PHRASE

AND WORK
ASSURE

4003 0 0 0 0 0 677 0 0 0 0 0
0 0 0 0 0 0 858 0 0 0 0 0

NOUN PHRASE

THE QUALITY
AND TIMELINESS

OF THE END PRODUCT

0 0 2007 0 0 0 5 0
4003 0 0 0 0 0 8888 0
0 3032 2007 0 0 0 620 0
0 0 0 8888 0 0 0 0

VERB PHRASE

0 0 0 0 0 0 0 0 0 0 0 0

VII. DISCUSSION

The semantic representation derived by the process described is now ready for further processing. Comparisons between documents can easily be made for looking for similar content, even if the documents use completely different wording. The programming of such content matching is relatively straightforward, because of the numerical array structures of the data. The AEG system uses such a technique when comparing student essays against model answers.

The AEG system based on this technique can process a 400 word essay in about 3 seconds.

The POS tagger, as with most taggers, does not accurately tag words in all cases, and so the chunking process does not produce completely accurate chunks. However this does not seem to hinder substantially the construction of a meaningful sentence summary.

The stemming program does not produce stems in many cases that are of the form required. Many of the stems are not real words, and so when the stems are used in the lookup process using the thesaurus, the words are not found. The stemming process needs to be modified to produce whole words, so that the success rate of stem lookups is improved.

VIII. REFERENCES

- [1] S.Alter, *Information Systems A Management Perspective*, Addison-Wesley, Reading, Mass.: 1992, 168.
- [2] M.Kolln, *Understanding English Grammar*, MacMillan, New York. 1994.
- [3] O.Mason, <http://web.bham.ac.uk/O.Mason/software/tagger/>
- [4] D.Perelman-Hall, "A Natural Solution", *Byte*, 17, 2, February, 1992, 237-244.
- [5] M.F.Porter, "An Algorithm For Suffix Stripping", *Program*, 14, 3, July, 1980, 130-137. (Code written by B.Frakes and C.Cox, 1986, and changed by C.Fox in 1990 and 1991).
- [6] M.Poesio, "Semantic Analysis", in R.Dale, H.Moisl and H.Somers, *Handbook of Natural Language Processing*, Marcel Dekker, New York: 2000.