View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by espace@Curtin

School of Information Systems
Curtin Business School

A Rubric Based Approach towards Automated Essay Grading – Focusing on High Level Content Issues and Ideas

Sean Lam Hon Wai

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University

August 2012

Declaration:

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

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

Signature: …….

Date: …….

Acknowledgments

Firstly, I would like to give honourable mention to the invaluable advice given to me by my supervisor, Professor Tharam Dillon, who has served as an outstanding supervisor in his support and understanding during the writing of this thesis. Also thanks to Professor Elizabeth Chang who, together with Professor Dillon for giving me the idea for this particular research and co-authoring several conference papers that relate to the thesis. The constant academic challenge and guidance by Professor Dillon is without question the motivation for the completion of this thesis. I would also like to thank Professor Elizabeth Chang for providing the conceptual input and also for the great research environment that was DEBII, Bruna Pomella for her many hours spent proof reading this thesis, Mr. Wei Liu for helping with the development of the system and Dr. Omar Hussain, whose help in the last few months of writing this thesis was more than I could have ever asked for.

Secondly, a special mention to Dr David McMeekin, for serving a short while as my co-supervisor, but whose contributions were absolutely incomparable near the end, and also for helping me from time to time by exchanging extremely important “data”.

I would also like to extend my thanks to friends and colleagues who have been with me throughout my time here. Valencia, for our many hours of conversations while brooding over coffee mugs, Farida, for always working at her desk making me feel guilty enough to continue working at mine, Davor, for our little coffee and occasional smoke breaks and of course, the biltong, Fedja for providing much needed advice when I was pretty stuck. There are many others who have had in some way made an impression on me during the course of this work, professional or casual and while I may not have mentioned them all, my heartfelt thanks and respect goes out to each and every one.

Thirdly, I would also like to thank my family for their constant support and empathy, my flatmate Partick (yes the spelling is intended) Fritz for just being there to share a quiet and sometimes not so quiet drink, together with the occasional cigar and smoked beef. My neighbour Shyrol, who is such a sweet girl for bringing me dinner and soup sometimes, and also warm lemon tea for those cold winter nights. Not forgetting the EG guys back in Singapore, Izzy, Casey and Kyla Foo, Hetal, Sera, Joel, Jamie, Sam, Lyn and many others for our many quirky moments and good laughs which I’ll always remember and CiCi for visiting me at my desk, always making me smile and laugh out loud, though most of the time at your expense.

Abstract

Assessment of a student's work is by no means an easy task. Even if the student response is in the form of multiple choice answers, manually marking those answer sheets is a task that most teachers regard as rather tedious. The development of an automated method to grade these essays was thus an inevitable step.

This thesis proposes a novel approach towards Automated Essay Grading through the use of various concepts found within the field of Narratology. Through a review of the literature, several methods in which essays are graded were identified together with some of the problems. Mainly, the issues and challenges that plague AEG systems were that those following the statistical approach needed a way to deal with more implicit features of free text, while other systems which did manage that were highly dependent on the type of student response, the systems having pre-knowledge pertaining to the subject domain in addition to requiring more computational power. It was also found that while narrative essays are one of the main methods in which a student might be able to showcase his/her mastery over the English language, no system thus far has attempted to incorporate narrative concepts into analysing these type of free text responses.

It was decided that the proposed solution would be centred on the detection of Events, which was in turn used to determine the score an essay receives under the criteria of Audience, Ideas, Character and Setting and Cohesion, as defined by the NAPLAN rubric. From the results gathered from experiments conducted on the four criteria mentioned above, it was concluded that the concept of detecting Events as they were within a narrative type story when applied to essay grading, does have a relation towards the score the essay receives. All experiments achieved an average F-measure score of 0.65 and above while exact agreement rates were no lower than 70%. Chi-squared and paired T-test values all indicated that there was insufficient evidence to show that there was any significant difference between the scores generated by the computer and those of the human markers.

List of Publications

(Sean) Hon Wai Lam, Tharam Dillon, Elizabeth Chang, Pornpit Wongthongtham, Wei Lui, "Detecting Events in Narrative Essays," Advanced Information Networking and Applications Workshops, International Conference on, pp. 363-368, 2012 26th International Conference on Advanced Information Networking and Applications Workshops, 2012

Lam, Hon Wai (Sean) and Dillon, Tharam and Chang, Elizabeth. 2011. Determining writing genre: towards a rubric-based approach to automated essay grading, in IEEE 24th International Conference on Advanced Information Networking and Applications (AINA), Mar 22-25 2011. Singapore: IEEE.

Lam, Hon Wai (Sean) and Dillon, Tharam and Chang, Elizabeth. 2010. Towards the use of Semi-structured Annotators for Automated Essay Grading, in Ismail, L. and Chang, E. and Karduck, A.P. (ed), IEEE international conference on digital ecosystems and technologies (DEST 2010), Apr 12 2010, pp. 228-233. Dubai, United Arab Emirates: IEEE.

Table of Contents

Part I: Introduction

List of Figures

List of Tables

List of Formulas

Chapter 1-Introduction

Assessment of a student's work is by no means an easy task. Even if the student response is in the form of multiple choice answers, manually marking those answer sheets is a task that most teachers regard as rather tedious. As Mason and Grove-Stephenson (2002) mentioned, a large amount of a teacher's time is spent grading students' work. After the inception of systems that were able to automatically grade multiple choice answer sheets, the next thought that predictably followed was "Could the same be done for essays?" Which in turn led to the next question "Is there a better way of automatically grading an essay?", ultimately setting the backbone of this thesis.

The second question has been partially answered over the last few decades, with the development of dozens of automated systems using various methods to provide a more convenient but no less effective means of grading (Larkey 1998, Perez-Marin 2009). Of course, no single solution is ever perfect, and thus, with each solution arose more questions and more problems. Questions such as "Will the system be better at grading than humans?" or "How can it be certain that the computer can understand language the same way that a person does?" Some critics have even stoked the fires of resistance by suggesting that these systems would take over the teachers' role, eventually rendering them obsolete and jobless.

However, there might, be some benefits that outweigh the problems. Human markers can be inconsistent and subjective at times due to certain judgements and biases and thus the same essay might have as many differing grades as it does markers (Streeter et

al. 2003). An essay might even be marked down simply because the handwriting wasn't as neat as the marker would have liked or even because the marker was having a particularly bad day. The time-saving factor is also seen as a big advantage over manual assessment, which usually translates into cost savings either in the form of opportunity costs by applying the time elsewhere, or by reducing labour costs since the grading process is automated (Chung and O'Neill 1997).

In this chapter, a brief introduction into the field of Automated Essay Grading (AEG) is given, followed by the other fields which play a part in this research such as Narrative Texts within the field of Narratology. The rest of this chapter details some on-going issues and challenges faced by AEG systems that gave rise to the motivations behind this research. Next, the current debate for and against the application and use of AEG systems is briefly presented. Finally, the scope and significance of this research project is discussed, followed by the outline and structure of this thesis.

Before we delve into the more intricate details of essay grading, however, there is first a need to have a better understanding of the different kinds of student responses that a teacher might encounter.

1.1 Essays and other types of Student Responses

An essay, a composition, an argument or an exposition: regardless of the name, in the context of a classroom it can simply be said that this type of response gives a student the chance to present his or her understanding of a specific topic or general subject domain. Robert Ebel stated that through an essay, an indication of the student's

thought process is shown, together with his/her ability to argue the reasons supporting a contention (Ebel 1979).

Though essays are more common in higher education (such as universities) as a form of pedagogical discourse, essay-type responses are often required of students at lower levels, including primary school. For example, it is not unusual for students aged between 13 and 16 years to be tasked with writing an essay about a certain subject, albeit with various prompts or leads. This is also true in Western Australia, where primary and junior secondary students in years 3, 5, 7 and 9 are asked to write an essay on a certain topic or subject as part of the state-wide literacy assessment under the Western Australian Literacy and Numeracy (WALNA) program.

An essay test in this case usually requires the students to write a short narrative essay, either based on a general subject domain or in response to various prompts or leads. Although Ebel describes a good essay as one where the student is able to draw from his or her own "command of an ample store of knowledge that enables them to relate facts and principles", these essays are not focused on a student's ability to articulate arguments or present facts in support of a claim. As opposed to the type of essays written at an undergraduate level, the main focus here is on the student's command of the English language as a whole (Ebel 1979).

Apart from essay tests, (Perez-Marin et al. 2009)) have listed some other forms of free text answers which they classified into different types of question groups based on three different criteria:

1. Number of correct answers
    a. Convergent Questions: only one correct answer. Focus on concrete fact. E.g. “Who was the first President of the United States?”
    b. Divergent Questions: many correct answers but are usually based on an opinion or hypothesis. E.g. “Which is the best way to conserve energy?”
2. Type of answers expected
    a. Open-ended Questions: many correct answers but are based on facts. E.g. “Describe the water cycle”
    b. Closed-ended Questions: where there is only one correct answer. Expected responses are ‘yes’ or ‘no’, ‘true’ or ‘false’ or a single keyword. E.g. “Is the Sun the only star in our solar system?”
    c. Counter-Questions: questions asked by a student for clarification. E.g. in response to the question “How would you measure the success rate of an organisation?” the student’s question might be “Which kind of organisation is in question?”
    d. Numerical Questions: where the answer requires some form of mathematical calculation
3. Function of the question
    a. Making a choice: usually in the form of multiple choices
    b. Determining if a sentence is true: the student is required to provide a response to a given prompt. E.g. “Having a large family is always better”.

The student might also be expected to provide justification for the claims for or against the statement

c. Develop Ideas: the student is expected to elaborate on a certain topic
d. Calculation: The answer is the result of numerical formulation and analysis

Though not an extensive list, the above questions cover most types of free text responses that are expected of students. The narrative type essays mentioned earlier might fall into the "Develop Ideas" category, although this does not encompass the entirety of the literacy assessment. Having gone through the various types of student responses, the next section provides a brief introduction to the field of Automated Essay Grading.

1.2 Automated Essay Grading

The vast amount of essays that teachers have to go through when marking has always been an issue; the task is relatively monotonous and time consuming, often taking up several hours which could have been better spent. This is not too great an issue if the number of students is relatively small, but the enormity of this task becomes exponentially more pronounced when the cohort comprises hundreds of students, as is usually the case in secondary and tertiary institutions (Chung and O'Neill 1997, Mason and Grove-Stephenson 2002).

Furthermore, the more time a teacher spends on marking, the less s/he has to spend on marking responsibly and conscientiously in order to provide a fair grade, rather than

simply going through the cursory motions of grading an essay based on a perfunctory glance. One of the first and most direct ways of alleviating this workload was to hire markers or extra staff to do the job; however, the costs of outsourcing this task often outweigh the benefits of the solution.

Additionally, even with the extra markers, marking standards among human graders are often inconsistent, giving rise to reliability and validity issues in the grades themselves (Streeter et al. 2003).

The development of an automated method to grade these essays was thus an inevitable step. According to researchers such as Valenti, Neri et al. (2003), interest in the development and application of AEG systems within the education community has increased over recent years, largely due to the increasing number of students attending universities and a growing interest in the on-line possibilities of such an application.

These systems have been described in many and various ways; some researchers refer to them as Automated Essay Grading (AEG) while others have called them Automated Essay Scoring (AES) or even Computer-Assisted Assessment (CAA). For the purposes of this thesis, automated systems shall henceforth be referred to as the first definition presented, Automated Essay Grading. Regardless of their names, AEG systems provide a method by which assessment of a student's work can be carried out automatically, with little to no human supervision required, thereby allowing teachers to better allocate their time elsewhere.

AEG systems have been defined as computer technologies that serve to evaluate and score the written prose, while at the same time stating that the purpose of developing such systems was to assist in “low-stakes” classroom assessments (Shermis and Barrera 2002; Shermis and Burstein 2003).

1.3 Challenges with Automated Essay Grading

Page’s Project Essay Grader developed in 1966 was the pioneer of automated grading systems, aimed at improving the grading process. With it, however, came a slew of issues, both internal and external. The former refers to the challenges the system itself faced; early automated essay grading (AEG) systems did not have the benefits of advanced computational linguistic tools that are available now.

Even during the early 2000s, the most advanced Natural Language Processing tools had not developed a more in-depth analysis of free text past performing Part of Speech tagging while limiting contextual analysis to simple phrases (Cheville 2004).

While technological advancements have produced text processing tools with greater computational abilities and hence, better methods of in-depth analysis, with the increase in analysis capacity comes an increased strain on available resources and thus increased costs. These costs are further amplified through maintaining analysis databases and keeping a level of consistency throughout.

In addition to the aforementioned issues, other general challenges that AEG systems face include:

- The grading process is centred on topical content only
- Any ambiguity in unstructured text cannot be handled
- Systems that take in key word occurrence as a primary score measure are severely disadvantaged since one word can be represented in several different ways
- Human markers might not be consistent in grading the same essay

Externally, AEG systems also faced several challenges related, in particular, to their acceptance by the general community, specifically the people in the education sector. These issues are described briefly in the following section.

1.4 The Current Debate

There have been many debates on the effectiveness of using a machine to grade an essay (Wang and Brown 2007), the most common being that a machine would never have the same cognitive capabilities of a human reader and would thus be unable to give a score that considers the more subtle aspects of written work. Building on these criticisms, other works by researchers such as Hamp-Lyons (2001), Chung and O'Neil (1997), Kukich (2000) and Rudner and Gagne (2001) have also stated other issues including:

- Lack of human interaction - some argue that having a computer grade an essay takes away the understanding of implicit meanings in text that only a human would be able to comprehend

- Susceptibility to being fooled by cheaters - based on the concept of some system's grading process being based on keyword identification or basic counting methods, some critics have stated that it would be relatively easy to fool the system into giving a better grade
- Need for a large training corpus - some systems require large amounts of manually graded essays to train on before being able to effectively grade an essay. This is especially true for systems that rely heavily upon the subject domain

The subject matter pertaining to the essay would inevitably lead to variations in the grades given depending on what one marker thinks is relevant and what another thinks is not. Valenti et al. (2003) have suggested that students may perceive this subjectivity as a source of unfairness.

1.5 Narratives

Narrative texts, in the context of narratology can be described as a piece of written work describing certain happenings from several perspectives. As Bal (1980) stated that if a text is a finite, structured whole composed of language signs, then a narrative text would be considered a text in which an agent relates a narrative; in this context, an agent could refer to a particular character within the narrative or the author him/herself.

In addition, Lucariello (1990) also stated what she felt were two essential characteristics of narratives stories. The first was the pentadic imbalance and the

second the consciousness or subjectivity of the protagonists. Pentadic imbalance refers to a skew from the normal, or as Lucariello stated, a “departure from the expectation or conventionality” (pg. 132).

An earlier work by Burke (1969) defined such a pentad. Stated by Burke as a minimum set of criteria, a narrative should at the very least contain an:

- Actor
- Action
- Goal or intention
- Scene
- Instrument

In relation to the second characteristic, also mentioned in previous works of Greimas and Courtes (1976) is the subjectivity of the protagonists which might also include the subjectivity of the narrator as well. Building on that, Lucariello took it to mean that any developed narrative should have a double landscape, where one of the worlds of action is described within the story and the other in the minds both of the protagonists and the narrator (Courtes 1976; Greimas 1989; Lucariello 1990).

Narratology can of course be related to more than just texts. Labov (1995), in the context of verbal communication described the narrative as “a method of recapitulating past experience by matching a verbal sequence of clauses to the sequence of events which actually occurred.” He goes on to explain that this concept of narrative could be further described as a “sequence of two clauses which are

temporally linked", and that a change in their order will result in a change in the temporal sequence of the original semantic interpretation (Labov 1995 pg. 360).

According to Labov, a fully developed natural narrative should consist of the following:

- Abstract
- Orientation
- Complication action
- Evaluation
- Result or resolution
- Coda

Although not specifically talking about narratives in the textual form, what Labov did mention were temporal clauses, which could be interpreted as one situation evolving into another based on past happenings. This is essentially what comprises a narrative.

1.6 Motivation

The main motivation behind this research is to determine whether, by combining concepts found in narratology and AEG technology, a different method of analysing free text might be developed which could provide a more effective means of essay grading. Moreover, among the several areas that AEG systems help to address, there are two other problem areas that motivate this PhD research:

- the need for a large amount of computational resources; and
- the dependency on a subject domain.

Thus far, the more successful systems rely on heavy computational methods in order to acquire a deeper understanding of free text. This might make for a much more cognitive marking process but it also means that the costs of using these systems become rather high. In a trial of the Intelligent Essay Assessor carried out by Williams (2004), the costs of using the system totalled over AUD 11,000. Following the mantra of "there has to be a better way", this thesis attempts to determine if a cheaper but no less effective method of essay grading can be developed.

The dependency of having pre-knowledge of the subject domain usually means that for every new subject domain that is being examined, the grading system would probably have to be trained again. This training process varies across different systems, some more tedious than others. Were this issue to be addressed successfully, it would definitely make a significant contribution to AEG technology.

Although there are many more areas that can be addressed in the field of essay grading systems, it is also important to define the boundaries of a research project, as in the case of this thesis. Therefore, it is essential to outline what will and will not be addressed within this work.

1.7 Thesis Scope

Even though the field of Automated Essay Grading can be said to be specific in itself, there are still a myriad of possibilities in the development of a particular system. For instance, the type of student response that the grading system takes as input would heavily influence how the system is constructed. For the purposes of this thesis, the

student response considered will be limited to narrative type essays written by students from WA, ranging from Years 1 to 12.

In addition, while the grading system is based primarily on the National Assessment Program – Literacy and Numeracy (NAPLAN) marking rubric, not all aspects of the rubric will be addressed in this thesis. The criteria from that rubric which this thesis attempts to address will be limited to those relating to the stylistic aspects of a narrative type essay.

One of the reasons why AEG systems are adopted is their perceived savings with regards to labour costs. To that end, in the course of this research project open-sourced tools are heavily utilised, thereby minimising the perceived costs of implementing the system. However, due to the time and resource limitations of this research, this thesis does not attempt to address in detail the financial benefits that may or may not be derived should the proposed solution be applied. This however, might be an avenue for future work.

1.8 Significance and Objectives

The contributing factors of this thesis span several areas. Firstly, there have not been any grading systems that operate outside a specific scope; most systems need to be trained on a specific domain to able to effectively grade an essay and this is especially crucial for systems that rely on keyword associations. Having a system that is able to perform a grading process regardless of the domain in question would be a significant step towards creating a more versatile grading process.

Secondly, oftentimes the ability of a system to effectively grade an essay at a higher cognitive level requires a large lexicon of terms or word senses. This is largely due to the vastness of the English language, thereby requiring a large knowledge base in order to handle all or at least most known word senses. This means that the system would be quite resource-intensive, requiring large amounts of computing power in order to operate at an effective level. After all, if the system is more costly and time-consuming than its human counterparts, there would be little reason to implement the system in the first place.

Finally, this thesis presents a novel method in which narrative essays could be looked at. By being able to identify specific parts of the essay that make up the essential parts of the story using basic NLP tools, it might allow a computer to have an understanding of the text through less resource-demanding methods. Furthermore, most essay grading systems are based on the respective developer's perception of how well an essay should perform; based on either predictive methods or pattern recognition type systems. Setting it apart from its peers, the proposed grading system is based on a specific marking rubric, tried and tested through its use in grading narrative essays.

The main objectives of this research are as follows:

Objective 1: To provide a comprehensive literature survey of works done in the field of Automated Essay Grading together with the methods upon which those systems are based.

Objective 2: To attempt to merge several theories regarding narrative essays with automated grading techniques and determine the outcome.

Objective 3: To design and develop an essay grading system that does not have to rely heavily on resource-intensive procedures, thereby minimising possible costs and allowing the system to run smoothly without excessive computation.

Objective 4: To test and evaluate the effectiveness of the grading system using actual student essays in an attempt to provide a proof of concept.

1.9 Thesis Structure

Including Chapter 1, this thesis is made up of 10 chapters, which are structured as follows:

Chapter 2: Provides a review of the work done in the field of Automated Essay Grading systems. The methods in which different systems carry out automated grading are examined together with their strength and weaknesses.

Chapter 3: This chapter discusses some of the problem areas within the Automated Essay Grading field together with the key definitions and terminologies that will be used throughout this thesis. From problems identified, the specific issues that this thesis will attempt to address are discussed in detail. These issues are then broken down into individual aims which are geared toward providing a solution for said issues. In addition, the NAPLAN marking rubric is discussed together with a brief description of

the marking criteria. The end of this chapter discusses the methodology which was applied within this thesis.

Chapter 4: The theoretical framework is presented in which an overview of the proposed solution is provided, with the solutions presented as a means to address the issues identified in Chapter 3. Details of how the proposed solution is formulated and designed are also covered in this chapter.

Chapter 5: The first part of the proposed solution, which is the Event Detection Process, is discussed in detail. In this chapter, the specific steps involved in this process are described, which eventually lead up to where it is possible to obtain and output that is used in the next stage of analysis. In addition, a discussion covering the results of experiments carried out the overall performance of the Event Detection Process is provided.

Chapters 6 through 9 provide details on the second stage of the solution.

In these chapters the marking criterion that are considered within this thesis are discussed. Each chapter provides a description of the respective criterion being considered and the hypothesis that it serves to substantiate. The results of the experiments conducted are then presented and discussed. The individual criteria that these chapters discuss are:

- Chapter 6 - Audience
- Chapter 7 - Ideas

- Chapter 8 – Character and Setting
- Chapter 9 - Cohesion

Chapter 10: The final chapter concludes this thesis by providing a recapitulation of the work that has been done and the outcomes arising from it. An overview of the limitations faced in the course of conducting this research is discussed, together with identifying the future work that could be undertaken to improve the solution proposed by this author.

1.10 Conclusion

While some may think that using AEG systems is an impersonal and potentially job destroying tool, it would be prudent to note that most researchers consider the development of AEG systems as a means to assist the teachers and not replace them (Mason and Grove-Stephenson 2002).

In the next chapter, an introduction to and discussion of Automated Essay Grading systems are given. The discussion provides details of the different methods the various systems have adopted towards automated essay grading and highlights the strength and weaknesses of each.

Chapter 2 - Literature Review

This chapter will cover the different styles and methods that have been developed for Automated Essay Grading systems, some of which are commercially available. The purpose here is to give an overview of the main techniques that are currently in use, followed by a more in depth description of the systems themselves. This section concludes with a discussion of the strengths and weaknesses of each of these systems, and includes a Table that summarises their features such as the methods used for essay grading and those aspects of an essay which are the main focus of each system.

2.1 Automated Essay Grading

The field of automated essay grading is relatively new and in its infancy, with a history of just over 40 years (Wang and Brown 2007). The main advantage of using an automated system to score essays is that the time spent by teachers on grading is significantly reduced, together with the fact that the system is unbiased and is likely to produce the same result for a similar essay, ensuring that the marking standard Is consistent. An automated system uses a standard marking rubric or scheme and eliminates the issue of subjectivity which characterises human markers.

Previously, essay grading systems could be divided into two main types: the first marks an essay according to its technical aspects such as spelling and grammar, the second considers the more abstract features of the essay. Page (1966) describes the distinction between the two as focusing on either content or style, with the former referring to

what the essay says and the latter referring to the "syntax and mechanics and diction and other aspects of the way it is said". More recently, grading systems have tried to incorporate both elements in their scoring mechanisms using various statistical or language processing methods.

2.2 Style

Following the description used by Page (1966), an essay's style could also be referred to as the technical features of the text such as the spelling, grammar and punctuation, and other language conventions. Usually, a good essay is characterised by correct grammar, a minimal number of spelling mistakes and a certain uniformity and consistency of format.

2.3 Content

The biggest challenge in essay grading thus far is that while it is relatively easy to evaluate the stylistic aspects of an essay such as grammar and spelling as described above, getting the system to understand the content features of the text is much more difficult. The problem is somewhat easier to tackle when the essay is written on a pre-assigned topic, thereby allowing the analysis to be contained within a certain domain.

Page (1995) stated that although it is not possible to measure the intrinsic features of a text by direct means, this can be done by finding possible correlations. For example, the fluency of an essay could be measured by the approximate correlational values or '*proxes*' between certain intrinsic aspects of the essay.

2.4 Approaches towards Automated Essay Grading

Of the earliest programs designed for automated scoring, the Project Essay Grader developed by Ellis Page used multiple linear regression to determine those weighted features of a text which were most relevant to that of a grade given by a marker. Those features were then used in turn to predict the score of an essay (Page 1966). Since then, there have been several other developments in essay scoring software that use a multitude of different techniques such as Latent Semantic Analysis, Natural Language Processing and Artificial Intelligence, to name a few (Landauer et al. 2003; Burstein et al. 2003; Rudner et al. 2006; Dikli 2006).

2.4.1 Statistical-based Methods

Regarding the use of statistical methodologies to determine an essay's grade, multiple linear regression is by far the most commonly adopted. Usually, this approach begins by identifying all possible features of an essay using various Natural Language Processing techniques, which a human grader might deem important enough to make a significant contribution to its score. Through a process of elimination, these features are gradually condensed to a finer set that includes only those features that make a large contribution to an essay's final score.

2.4.1.1 Project Essay Grader (PEG)

As mentioned above, the Project Essay Grader developed in 1996 could be considered as the pioneer of today's essay grading systems. Page proposed that it is possible to identify which features of a passage have the most influence on the score that a human

marker would give; once those features have been identified, multiple regression is used to compute a predictive formula for scoring an essay. 'Trins' (in**trin**sic) are those intrinsic aspects of an essay (e.g. fluency, diction, grammar, punctuation, etc) which Page determined to have a high weighting according to a human grader while 'Proxes' (ap***prox***imated) refer to the correlation of those intrinsic variables (Page 1966; Wang and Brown 2007).

The scoring stage uses the two main variables, *Trins* and *Proxes*, gathered in the training stage from a test sample of 100-300 training essays to predict the score of an unmarked essay, with the final score depending mainly on the linguistic aspects and style of an essay as evaluated by the PEG system (Page 1966; Williams 2001; Dikli 2006). An evaluation conducted by Page himself using about 30 Proxes produced promising results, with the correlation between the PEG system and human graders at .78 although this varies in later evaluations (Kukich 2000; Williams 2001).

A strength of the PEG system is the reasonably high correlation between human grader scores and the system-generated scores (some reaching as high as 0.85 between two or more graders); another is that the system is able to track errors, allowing for greater ease of evaluation (Kukich 2000; Chung and O'Neil 1997).

Having said this, the weaknesses of the system are that since the contextual features of the essay such as organisation are not detected, constructive feedback is not given. Furthermore, with only a surface scrutiny of the features, it is entirely possible to trick the system into giving a higher score by writing a longer essay without contextual

reference to the topic (Dikli 2006; Kukich 2000). Since the 1990s, PEG has undergone several modifications whereby several lexicons were combined with specific parsers.

2.4.1.2 MarkIt

A more recent essay grading tool, MarkIt , was proposed by Williams and Dreher (2004) which made use of a rough clustering or "chunking" of the text in order to obtain sentence structure, represented by Noun Phrases and Verb Clauses which relate to the context and actions pertaining to the subject respectively.

According to the developers, Verb Phrases are extremely complex, thus prompting them to use Verb Clauses together with Noun Phrases. By mapping the root meaning of the word to the one found in the text, thereby assigning it the thesaurus index number, a numerical representation of the text can then be established. These are then used in a classification approach of predicting an essay's score using multiple linear regressions, with vector space computations formulating some of the calculation inputs.

Some of the issues arising from this method are that the system seems to simply use a version of Named Entity Recognition, where Noun and Verb Clauses are identified and counted, similar to the Bag of Words approach. While MarkIt might produce an accurate score with a high agreement rate among human graders under some circumstances, it would be easy to trick the system into giving a high grade if the mechanics of the algorithm are known even generally (e.g. including more keywords to attain a higher Noun Phrase value).

Furthermore, it appears that the system is unable to handle word sense disambiguation. A short in-depth evaluation conducted by the authors showed small inconsistencies between human graders and the IEA system, although there were some cases with larger differences (Williams and Dreher 2004).

2.4.1.3 Latent Semantic Analysis

The fundamental logic of Latent Semantic Analysis (sometimes known as Latent Semantic Indexing) is that the meaning of a body of text is dependent on the meaning of each and every one of the words used and the modification of any word would affect the meaning of the passage in one way or another (Dikli 2006). Hence, it can be said that LSA represents the meaning of a word as an average of all its meanings in the passage in which it appears and similarly, the meaning of a passage as an average of the meaning of all the words within (Landauer et al. 1998).

As described by Landauer et al. (2003 p.88), "meaning of word 1 + meaning of word 2 + ... + meaning of word n = meaning of passage". Therefore, this makes it possible for passages that contain different words to have the same meaning and vice versa.

Foltz (1996 p.198) described LSA as a "statistical model of word usage that permits comparisons of the semantic similarity between pieces of textual information". In the first stage of this approach, a term document or occurrence matrix is constructed to represent how many times a term appears within a body of text, with each row representing a unique word while the columns refer to the context in which it is used.

The second stage involves applying SVD to the matrix, whereby it is broken down into three separate matrixes the product of which would once again be the original matrix.

Landauer et al. (2000) explains that LSA analyses the semantic relations between a set of textual documents and the terms within it through a series of concepts contained in a general body of text; this gives the impression that LSA performs contextual analysis, which is not always the case.

In fact, it is the co-occurrences that enable the formation of groups or clusters of related concepts, although just because there is a co-occurrence between two concepts, this does not necessarily mean that they share the same context. On the contrary, it might be entirely possible that the two concepts refer to the co-occurring concept in totally different contexts; this is especially so in long documents (Garcia 2007).

2.4.1.4 Intelligent Essay Assessor System (IEA)

Mathematically speaking, the system uses a technique known as Singular Value Decomposition (SVD) and LSA is an application derived from it. Developed by the University of Colorado and purchased by Pearson Knowledge Technologies (PKT), the Intelligent Essay Assessor's (IEA) main schema is founded on Latent Semantic Analysis (LSA). The system places more emphasis on the context of the text rather than using the common approach of scoring based on formal aspects such as grammar and punctuation although these are also incorporated into the scoring model (Dikli 2006; Williams 2001). An overall view of the system's architecture is shown in Figure 2.1.

The system operates in a manner similar to Page's essay grading system, PEG, in that it tries to pick out certain semantic features of a good essay then assesses an unmarked in terms of those features. The first stage involves training the program on a large background of a particular domain, be it from essays pre-scored by human experts, textbooks or other sources of semantic variance, in order to "establish a semantic space" (Yang et al. 2002 p.395) for the domain. The next stage involves comparing those semantic features or concepts to an unmarked essay to predict the score. By mapping a student's essay to the training set, the system is able to identify a range of semantic possibilities within the essay, thereby generating a holistic score or feedback based on the level of similarity (Foltz et al. 2000; Yang et al. 2002; Warschauer and Ware 2006).

One of the advantages stated by Kukich (2000) and reiterated by Wang and Brown (2007) is that the system is able to "capture transitivity relations and collocation effects among vocabulary terms, thereby letting it accurately judge the semantic relatedness of two documents regardless of their vocabulary overlap" Kukich (2000, p.24-25). Above all, what makes IEA stand apart from other current systems is its ability to detect plagiarism, which escapes most human markers since it is a tedious task to perform, especially when a large number of essays are to be graded. A survey conducted by Williams (2001) reinforces the abovementioned points when 327 essays were sent for grading by IEA. The system managed to detect a few cases of plagiarism that had escaped the notice of human graders (Williams 2001).

A prominent issue with the IEA is the number of essays required for training (roughly in the vicinity of 100-300). Even PKT, the producers of the system, concede that this issue needs to be addressed in order to improve the system, although other systems have an even higher number required (upwards of 300). Another issue is that since the system depends partially on essays graded by human experts, the costs of training the system might be more than some organizations are able to afford, particularly since this is in addition to the high computational costs of LSA. Furthermore, for all the analysis on content that the system performs, creativity as well as critical and reflective thinking by the student is not taken into account when calculating the essay score (Dikli 2006; Landauer et al. 2003).

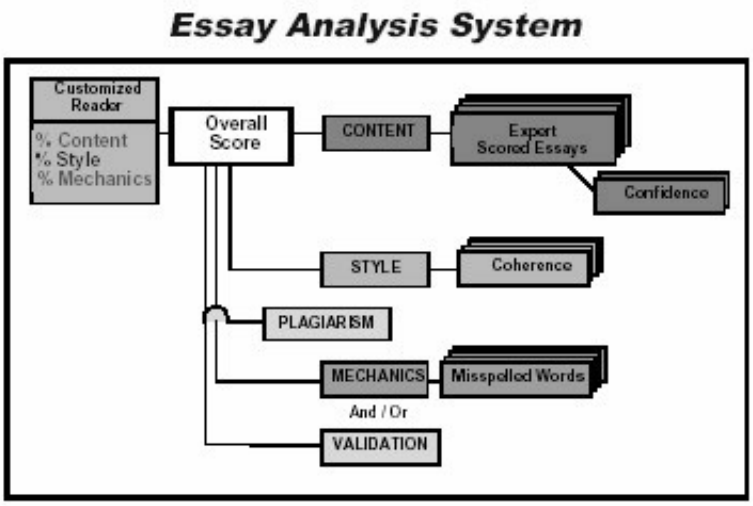

Figure 2.1: IEA System Architecture, source Knowledge Analysis Technologies

2.4.2 Bayesian-theory-based Approaches

The Bayesian or Bayes' conditional probability theorem is a common method of measuring probabilities regardless of the interpretation placed on the values used in

the prediction algorithm. For a more detailed explanation of Bayesian theory approaches we refer to Bernado (2000).

The basic principle here is to determine the probability that a result is true, with the knowledge that another result is true. This theory differs slightly from the conventional Bayes' theorem in that the values that make up the algorithm are subject to certain probabilities. In applying this theory, one would first have to specify the likelihood of a hypothesis being true, the value of which is then manipulated by the discovery and inclusion of relevant data (Berger 1985).

A simple explanation of the Bayes' conditional probability theorem is as follows: say, for example, that we have hypothesis *H*, which can be a statement that is believed true or a certain numerical value. The probability that *H* is true before considering all other forms of data that might affect the outcome is defined as the *prior* probability. Conversely, the posterior probability is the likelihood that *H* is true *after* the discovery or relevant data. This is further affected by the conditional probability of actually discovering new data should *H* hold true. The probability of discovering new evidence even with *H* being null is also taken into consideration. Therefore, the main objective is to determine the posterior probability, knowing the values of the other variables. This can be represented by the formula:

$$\mathrm{P}(H|N) = \frac{P(\mathrm{N}|\mathrm{H})\mathrm{P}(\mathrm{H})}{P(N)}$$

Formula 1: Baye's conditional probability

Where:

- H = original hypothesis
- N = new data and evidence that can be observed
- P(*H*) = prior probability
- P(N|*H*) = conditional probability of seeing N if H happens is true
- P(N) = probability of E under any circumstances
- *P*(*H*|N) = *posterior probability*

There are two more well-known AEG systems that incorporate Bayesian theory into their marking schemes, the Bayesian Essay Test Scoring System and the Text Categorisation Technique. The former uses Bayes' conditional probability theory as the underlying principle of predicting essay scores while the latter includes other statistical methods apart from the Bayes' theory.

2.4.2.1 Bayesian Essay Test Scoring System (BETSY)

Using Bayesian theory as the underlying approach, the Bayesian Essay Test Scoring System (BETSY) was developed by Lawrence M. Rudner and is open sourced. According to Rudner and Liang (2002), the BETSY scoring approach can be seen as an extension of Bayesian Computer Adaptive Testing (CAT), the extension being that of classifying the text according to a four-point nominal scale (e.g. extensive, essential, partial, unsatisfactory), using a large set of items. Here, they refer to items as "a broad set of essay features including content features (specific words, phrases), and other essay

characteristics such as the order certain concepts are presented and the occurrence of specific noun-verb pairs" (Rudner and Liang 2002 p.4).

The system utilises text classification in the form of two models, the Multivariate Bernoulli Model and the Multinomial Model. Generally speaking, while the former takes in each essay as a "special case of calibrated features" (Dikli 2006 p.20) and checks whether or not these can be found in an essay, the latter views essays as an example of those features and checks the number of occurrences of these features within the essay (Rudner and Liang 2002; Valenti et al. 2003; Dikli 2006). In addition, the system also incorporates NLP features into the scoring model, allowing users the option to include stemming and identifying stop words which might improve system performance and exclude erroneous results respectively, and also a form of feature selection based on entropy to improve the system's accuracy (Rudner and Liang 2002).

2.4.2.2 Text Categorization Technique (TCT)

Developed by Leah S. Larkey in 1998, the Text Categorization Technique (TCT) uses an approach based on distinguishing the "good" essays from the "bad" using output from trained Bayesian classifiers and other techniques to grade essays (Larkey 1998; Williams 2001). Larkey conducted experiments using five data sets that had been previously graded by hand; the subjects of the training essays used were social studies, physics and law and two question sets. The first required a student to present an argument and the second asked the student to evaluate an argument. According to Larkey (1998), only the first question set would be evaluated more according to the cohesive and logical flow to the text, rather the mention of key points.

To evaluate the system, she conducted two separate experiments consisting of several techniques. The first experiment used three datasets that consisted of social studies, physics and law essays and was conducted by training Bayesian and k-nearest neighbour classifiers. The performances of both methods were then compared by using linear regression (Larkey 1998).

The Bayesian classifiers were trained to distinguish between good and bad essays, with the data sets divided at certain points. For example, if the scale used to determine good to bad essays were a four-point scale (1-4), separate classifiers would be trained to distinguish those essays in the first category from the rest, and so on.

Text complexity features used in linear regression started with the removal of 418 stopwords, with the remaining words stemmed using the k-stem stemming algorithm. Candidate features were identified as words that appeared in at least three essays. Candidate features were then filtered into a set that had the highest correlation with manually-assigned scores.

Using the K-nearest-neighbour classifiers, training set essays that were most similar to the test essays were identified, giving the test essay a score similarly weighted to the average of the k essay group.

The text complexity feature set identified by Larkey consisted of eleven features namely:

- Chars – number of characters in the essay

- Words – number of words in the essay
- Diffwds – number of unique words
- Rootwds – forth root of Words value
- Sents – number of sentences
- Wordlen – average word length
- Sentlen – average sentence length
- BW5 – number of words longer than 5 characters
- BW6 – longer that 6 characters
- BW7 – longer than 7 characters
- BW8 – longer than 8 characters

Using three separate variable sets, which were a mixture of the text complexity features, Bayesian classifiers and k-nearest neighbour score, linear regression was performed to see which variables were attributed to the highest variance within the data and their coefficients, after which an essay score prediction equation was derived from those variables.

Results from the experiment showed good performance on the social studies and law datasets while a poorer performance was observed in the physics dataset with exact agreement rates ranging from 0.50 to .65 (Larkey 1998). The second experiment was carried out using the next two data sets, with exact agreements ranging from 0.65 to 0.88.

Judging from the correlation scores, the system performed reasonably well on some datasets while exceeding expectations on others. In the evaluation of the text categorization technique, Larkey stated that the k-nearest-neighbour approach performed much poorer than the other two approaches, although she suggested that applying more sophisticated features or a different similarity metric might improve performance.

2.4.3 Natural Language Processing

Of the many types of Human Language Technologies (HLT), Natural Language Processing (NLP) is probably the most complex. Uniquely distinguished from other forms of HLTs such as Text Mining, Summarization or Generation, NLP is the practice of understanding the content of the text, rather than focusing mainly on extracting key pieces of information such as in Text Mining.

2.4.3.1 Electronic Essay Rater (E-Rater)

Of the few AEG systems that consider the linguistic features of a passage, E-Rater developed by the Educational Testing Service, is one. Originally, the system was called Computer Analysis of Essay Content and was used for grading the Analytical Writing Assessment part of the Graduate Management Admissions Test (GMAT), although since January 2006, it has been replaced by the IntelliMetric grading system. The E-Rater produces a holistic score on a scale of 0 to 6, and if the difference between the automated score, when compared to the human rated score is more than 1, another human grader is used to settle the discrepancy (Yang et al. 2002; Dikli 2006).

Three specific modules of the system are used to identify certain characteristics of an essay such as the syntactic module, in which a parser is used to "identify [ies] syntactic structures, such as subjunctive auxiliary verbs and a variety of clausal structures such as complement, infinitive, and subordinate clauses" (Burstein et al. 2003 p.1) to pick out syntactic variety; the discourse module, in which a conceptual framework based on relations between conjunctions such as cue words (e.g. "probable" or "likely" to express a chance or probability), terms which could be in the form of conjuncts ("to summarise" or "to conclude" when summarising a passage) and syntactic structures, is used to consider the organisation and structure; finally, the topical analysis module picks out topical content and variety in the vocabulary (Burstein 2003; Burstein et al. 2003; Burstein and Marcu 2000). The topic analysis module employs a technique known as vector-space modelling whereby, as described by Burstein (2003 p. 117), "training essays are converted into vectors of word frequencies, and the frequencies are then transformed into word weights". Similar to the one provided by Dikli (2006), Figure 2.2 further illustrates the transformation of training essays into weight vectors:

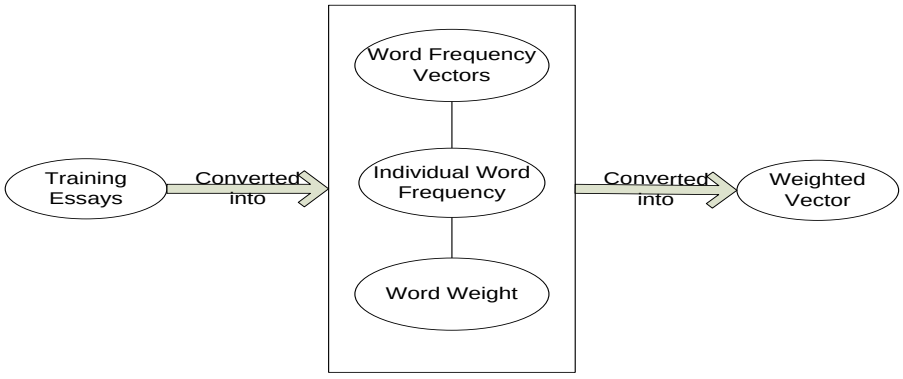

Figure 2.2: Weight Vector Transformation

Having identified the weighted features which make up a good essay, the E-Rater then compares every new essay that it evaluates against those features, using step-wise linear regression to create a scoring model that best predicts the score an expert human rater would give (Kukich 2000; Yang et al. 2002).

Summarizing the above, the E-Rater system incorporates a combination of Natural Language Processing (NLP) techniques and statistical techniques which are used to pick out specific features from a test bed of sample essays which then provide the basis of the scoring model (Williams 2001; Burstein 2003). The general assumptions or axioms of the E-Rater system are one good essay would not be that much different from another good essay and likewise for poor essays.

While the system as evaluated by Burstein and others in 1998 found the agreement rating between the system and human graders to be as high as 94% (Burstein et al. 1998), the fact remains that the system does not actually perform an analysis of the text since the scoring model is derived from the sample essays and every new essay is graded against it. Even though the system incorporates a set of more than 60 features (Attali and Burstein 2006), Powers et al. (2002 p. 116) stated in an evaluation of the E-Rater that it is not yet ready to function without human intervention, which is required to "keep E-Rater from seriously mis-scoring some essays."

2.4.3.2 E-Rater V.2

The mechanics of the E-Rater V.2 scoring system remains largely similar but improves on its predecessor by significantly reducing the number of features, condensing them

into a smaller set of more meaningful features which include Grammar, Style Measures, Organisation, Lexical Complexity and Prompt-specific Vocabulary Usage (Attali and Burstein 2006). The other improvement is that it allows for a greater degree of standardisation since it can create a single scoring model from the feature set. However, the issues mentioned previously are still present; while the feature module 'Lexical Complexity' considers word-based characteristics, key word frequency and word length do not necessarily measure the creativity of the writer per se.

2.4.3.3 Conceptual Rater (C-Rater)

Also developed by Burstein and others, the C-Rater uses many of the techniques of the E-Rater, the main difference being that the former was aimed at grading short-answer responses. The question types were similar to the short exercises commonly found at the end of textbook chapters (Burstein et al. 2001).

While many of the techniques used are similar, C-Rater focuses on content rather than style. The system also does not assign a holistic score; rather, in determining a right from wrong answer, the C-Rater searches for specific concepts within the given response. An advantage of using the E-Rater method for short answer responses is that a smaller training set can be used. Spelling mistakes, syntactic and inflectional variations together with semantic word senses also do not have much impact on the scoring system.

C-Rater was tested on a university virtual learning program and achieved an agreement rate of over 80% with a human grader. In addition, Leacock (2003) added that when

used in large-scale assessments which included roughly 100,000 short-answer responses, 19 comprehension and 5 algebra questions, the system attained an accuracy rate of 85%.

The shortcomings of the C-Rater are that, due to the stemming phase of the marking process, any answer that is dependent on verb tense is not assessable. Moreover, answers that include a quotation also seem to cause a problem while expressions that are not commonly used cause some confusion.

2.4.3.4 Schema Extract Analyse and Report (SEAR)

The Schema Extract Analyse and Report (SEAR) system was created in 1999 by Christie during the course of his PhD research. The system takes in word-processed essays as input and assesses essays on both style and content.

Four main stages make up the scoring method for SEAR, namely the Schema, Extract, Assess and Report stages. Firstly, the Schema stage sets up the marking criteria; for content, a model essay has to be prepared by the examiner beforehand in regards to the content that the student is expected to cover. In terms of style, the system is fed with weighted features to look for in a student essay which is then compared against that feature set. Secondly, the Extract stage utilises separate software that pre-processes the essays, although the same process is used for the style and content marking components.

Next, the Assess stage, used to mark an essay for style and content, is carried out using separate softwares. Both software components have to be run one after the other

should the marker wish to assess style and content together. The content schema prepared in the Schema stage is converted into a computer file which is then used to grade the essay's content, which is assessed by matching what the student has written with what was prepared by the marker in the Schema stage. (Christie 2003). The grading process is based on keyword matching and the relationship between those words. Each keyword is assigned a weighted score and if a student mentions those words in the correct relation, the score is achieved.

Lastly, in the Report stage, the results from the Assess stage are viewed. According to Christie, this stage is not involved in the actual grading process but instead allows the results to be viewed using a variety of preferred text editors. In addition, it also allows for the conversion of the result file into other formats for further analysis.

While the style assessment function was not field-tested at the time of publication, an evaluation of the system showed a human to computer Spearman correlation coefficient of 0.596 at its highest. Christie himself noted that while there may be a statistical significance in some cases of human to computer agreement, the SEAR system would perform well only on essays that are heavy on facts. Furthermore, the system would perform poorer should the volume of content and marks allocated in the marking schema increase (Christie 2003).The system also suffered from confusion due to spelling and grammar mistakes, and also from the variety of expressions used to convey the same point.

2.4.3.5 Automark

Developed in the UK by Mitchell and others in 1999, the Automark grading system was aimed at marking short ,open-ended responses and is now commercially available. The system utilised NLP techniques to “perform an intelligent search of free text responses for predefined computerized make scheme answers” (Mitchell et al. 2002 pg. 235-236). An overview of the modules comprising Automark is given in Figure 2.3.

Automark processes student responses via the following steps: firstly, templates are created by human experts for answers that are acceptable and answers that are not. Those answers are defined by the presence of certain parts of speech such as nouns, verb and prepositions. In the second step, the system takes into account the technical aspects of the text such as spelling and grammar. The third step involves sentence parsing to identify the main syntactic elements within the student’s response; this step also involves information extraction that is used to perform a high level extraction of specific concepts.

In the next step, a pattern-matching process is used to determine if those syntactic elements found in the student’s response match those within the predefined answer template created in the first step. Finally, the feedback module processes the output of the previous step to provide feedback to the student.

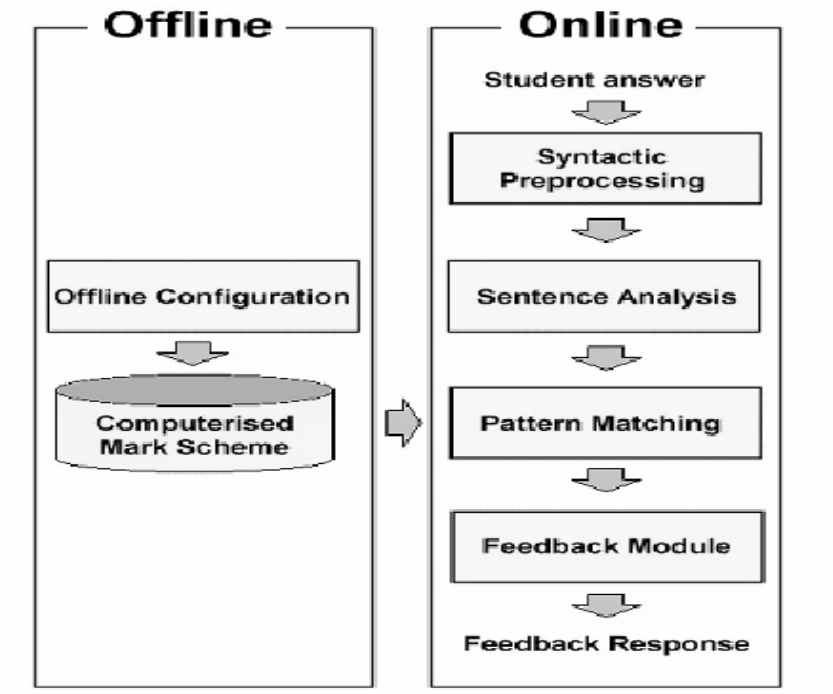

Figure 2.3: Overview of Automark, Source Mitchell at al. (2002)

Automark was used to assess a set of responses from the science curriculum of a class of 11-year-old UK students. Four types of questions were used with expected responses in the form of single words or values, short sentences and a description of data patterns. An evaluation showed a relatively high accuracy rate of 93% when comparing computer scoring with human marker scores, increasing to 96% when using a revised scoring template.

While the Automark system is able to ignore errors in spelling, typing and other features of free text that do not hinder the comprehensibility of the responses (Cotos and Pendar 2008),it is unable to effectively cope with spelling errors and poor sentence structure that is found within a correct answer, causing confusion within the system.

Other limitations also include problems in identifying incorrect answers and assessing more complex answer structures that were not specified in the marking schema.

2.4.3.6 PS-ME

Offering summative and formative assessments, the Paperless School Free Text Marking Engine (PS-ME) was developed by Mason & Grove-Stephenson in 2002. Primarily developed for scoring low stake and short-answer essays, the system is now commercialised and employed by some publishers. The system made use of several NLP techniques to analyse aspects such as grammar, contextual meaning and response-to-model answer comparisons.

Founded on the principles of Bloom's Taxonomy of the cognitive skills of knowledge (Bloom 1956), PS-ME comprised 3 subsets:

- Knowledge level: Much the same as Bloom's knowledge competence. According to the system developers, only the most relevant concepts need to be detected in a student's essay to evaluate his or her knowledge of the given subject
- Understanding level: Little detail was provided regarding this level due to its commercial sensitivity. All that can be said is that this stage comprises processes using the comprehension, application, analysis and synthesis levels of the Bloom Taxonomy as its foundation (Mason and Grove-Stephenson 2002)
- Evaluation level: Derived from the evaluation competence of the Bloom Taxonomy, the evaluation is based on the frequency of adjectives and adverbs, together with identifying certain syntactic patterns.

The training phase of the system required a minimum of 30 hand-marked essays which also included a number of poorly scored essays, used as a negative example. Essay scores were given upon the comparison of the essay against the model answers through the use of regression techniques. Feedback was also given in the form of automatically selected comments, thus allowing limited but formative feedback to be given based on different areas of the subject domain.

One of the system's limitations, according to Mason and Grove-Stephenson, is that essay grading could not be carried out in real time due to process requirements. Instead, essays were converted to XML files which were then sent to a web-based queuing system. One of the main drawbacks is that the system cannot cope well with spelling and grammatical mistakes and the selection of appropriate master texts. Calibration of the scoring process was also difficult due to the variability of human marker agreement.

2.4.4 Artificial Intelligence

2.4.4.1 IntelliMetric

Probably one of the first to utilize Artificial Intelligence (AI) into its scoring model, IntelliMetric developed by Vantage Learning between 1997 and 1998 is used widely across the United States (Rudner et al. 2006). While many details of IntelliMetric remain a closely guarded secret by Vantage Learning, the general architecture uses a mixture of AI, Natural Language Processing (NLP) and Statistical tools. IntelliMetric differs slightly from the more common approach of other automated grading systems;

instead of specifying a set of features before training the system on a data set, the system instead generates a scoring model from a set of marked essays prior to specifying a set of features or rubrics (Yang et al. 2002).

Broadly speaking, the first stage involves the previously mentioned step of analysing a training set consisting of essays pre-scored by human experts. In the next step, the scoring model is built by identifying the characteristics of essays at different score levels. Having done that, another set of training essays are run through the model to test its validity and effectiveness. The last stage is where the system generates the score of an essay by applying the scoring model to an unmarked essay (Wang and Brown 2007). Using this method, the system is said to mimic the way expert human-raters grade an essay by picking out those features or characteristics that they believe make up a good essay and those that do not.

The final score of an essay assessed by the system is based on the analysis of 72 different features, categorised into five groups of Latent Semantic Dimensions (LSD) (Elliot and Mikulas 2004; Dikli 2006; Ben-Simon and Bennett 2007) namely:

- Focus and Unity – attributed to cohesiveness and consistency in the writers' focus on the main idea
- Development and elaboration – relates to the expansiveness of content and support for arguments
- Organization and Structure – measures discourse logic and transitional fluidity within the passage

- Sentence Structure – complexity of a sentence including language use, readability and syntactic variety
- Mechanics and conventions –relates to adherence to standard English language rules (e.g. grammar, spelling, etc)

In addition, Dikli (2006) lists five underlying principles of the IntelliMetric system:

- Modelled on the human brain – the system is designed to mimic the way a human scorer or reader would process information it sees, which creates a sort of 'neuro-synthetic' logic processing which is said to mimic the way a human would think or, in other words, is 'brain-based' (Elliot and Mikulas 2004; Dikli 2006).
- The system is able to 'learn' – IntelliMetric can be seen as a learning engine in which useful information is acquired through 'learning' the way human experts carry out the scoring process. Essentially, the system is able to pick up how those characteristics human experts value in a good essay are identified.
- The system incorporates a complex step-based information processing system.
- Inductive reasoning - there have been suggestions that IntelliMetric is inductive, in that it is able to utilize inductive reasoning to make judgements on how the text is analysed based on the essays pre-scored by human experts.
- Multidimensional – instead of the common linear modelling approach, the system utilises several mathematical models a follows a non-linear, multi-dimensional approach when scoring an essay.

The tools used to develop the system are Vantage Learning's CogniSearch and Quantum Reasoning tools and technologies (Dikli 2006). The former is a tool specifically designed and created for use by IntelliMetric, allowing the system to incorporate NLP into the scoring process. This is achieved by "parse[ing] the text to analyse the parts of speech and their syntactical relations with one another" (Dikli 2006 p.15).

As found by an evaluation conducted by Rudner et al. (2006), the IntelliMetric system was able to closely match the scores given by human graders, with the only small issue being that the system tended to give slightly higher scores, but a further investigation on the researchers' part concluded that the issue is possibly insignificant since scores given by both human graders and the system fluctuated either way. Overall, the evaluation was extremely favourable to the IntelliMetric system. Another attribute of IntelliMetric worth mentioning is its ability to evaluate essays written in languages other than English including Spanish, Hebrew, Dutch and French (Elliot 2003).

2.4.5 Neural Network and Semantic based Systems

Inspired by the workings of the human brain, the basic logic behind neural-network-based-systems is to utilise artificial networks to learn different characteristics or features of an essay, which are then incorporated into the scoring model. Artificial neural networks consist of groups of neurons or nodes that are interconnected through non-linear computational or mathematical logics. The result of this is that the system is potentially able to identify complex relationships between nodes and in the case of essay grading, those nodes can represent individual essays in relation to their scores, or

even different features in those essays and how they each contribute to the final score. The application of an artificial neural network to essay grading might also produce deeper context analysis, allowing a system to 'understand' an essay through the relationships between words.

Using a different approach to achieve content understanding, semantic networks employ various forms of logical inference through a knowledge base. These knowledge bases themselves can take the form of word taxonomies, concept hierarchies or ontologies which can be used to define and eventually navigate through the different relationships between specified concepts or words.

Two systems could be found in the literature that use neural and semantic networks respectively, as the basis for their scoring models: the Intelligent Essay Marking System and the SA Grader.

2.4.5.1 Intelligent Essay Marking System (IEMS)

Developed by Ming and others at Ngee Ann Polytechnic, the IEMS focused on content analysis of short, qualitative essays. The system is based on Indextron, defined as a specific clustering algorithm, implemented with a Pattern Indexing Neural Network (Ming, Mikhailov and Kuan 2000). Though in itself not a neural network, Indextron as a clustering algorithm could be implemented as a neural network. The grading process is carried out by performing pattern recognition using some NLP techniques on the essays, wherein the patterns refer to the words.

The performance evaluation was carried out in a test that involved 85 third-year students of Mechanical Engineering. Students were asked to write a summary in not more than 180 words on a passage about cyberspace crime. IEMS performed relatively well, achieving a correlation of 0.8; furthermore, immediate feedback was available to students after submitting their responses.

2.4.5.2 SAGrader

The SAGrader is a commercially available system developed to assess essays based on their content. The system makes use of semantic networks comprised of knowledge of specific domains according to the essay topic. Networks were manually constructed and were made up of concepts and the relations between each of them, together with the features of these concepts, which were in turn used to determine if those same features could be found within a student essay.

Due to the domain-dependent nature of the network, the SAGrader is more suited to subject domains in which there are a limited number of possible responses, thus making it unsuitable for essays such as creative writing that have more subtle or tacit aspects. The system, like many others, is unable to process concepts that either contain misspelt words and/or are expressed in a manner not recognised by the semantic network.

2.5 Problems faced by current systems

The general trend of most systems developed for automated scoring is to follow the statistics-based, semi-supervised machine learning direction, in that training data (the

number of which varies between systems), is required for the system to learn either the model answer or a set of criteria used to grade the essay.

The other popular approach taken is to adopt NLP techniques together with those mentioned above to analyse the more tacit features of free text, each with varying degrees of success.

The main problems that seem to plague current AEG systems are that most of them are unable to cope with incorrect grammar or misspelt words in student responses. In such instances, systems that use a keyword detection method would be most affected since the absence of a particular keyword or concept would reduce the final score; even if the intention to express that idea were present, just an incorrect spelling or grammatical error would cause the system to overlook it.

While the answer should be penalised due to those errors, most systems found in the literature would treat the keyword or concept as an omission, which is unfair to the student. Valenti et al. (2003), in addition to the problem mentioned above, also mention other problems that AEG systems suffer from, namely the inability to:

- handle sentence structure correctly;
- identify an incorrect qualification; and
- provide mark scheme template.

The authors propose that the adoption of NLP techniques might reduce the impact of these problems since the use of sentence analysers, spell checkers and semantic processes would provide a better analysis of the text.

However, it might also be said that systems which depend heavily on NLP methods would be most affected if there were no measures to effectively parse textual content whilst handling spelling errors. An inability to perform such a step would pose a big problem when grading student essays from a lower level since there are many mistakes throughout.

Table 2.1 gives a summary of previous works, sorted according to the year they were developed. The performance column is based on the results stated by the developers, according to the type of performance measures that they employed. While some developers use the agreement rate between human markers and the system as a performance measure, others rely on correlation and accuracy scores while only Larkey's system of the text categorization technique published an exact agreement rate.

On average, those systems using the agreement rate as a performance measure achieved a score of 0.90, while those using the correlation measure achieved an average of 0.70. C-Rater and Betsy, using system accuracy as a performance measure, achieved an average of 0.91.

AEG System	Developer	Year	Technique	Performance	Main Focus
Project Essay Grader (PEG)	Ellis Page	1966	Multiple Linear Regression	Correlation of 0.87	Style
Intelligent Essay Assessor (IEA)	Landauer, Foltz and Laham	1998	Latent Semantic Analysis	Agreement rate of 0.85	Content
E-rater	Jill Burstein	1998	Natural Language Processing Statistical-Computation	Agreement rate of 0.87	Style and Content
IntelliMetric	Scott Elliot	1998	Artificial Intelligence Statistical-Computation	Agreement rate of 0.98	Style and Content
Text Categorization Technique	Leah S. Larkey	1998	Bayesian Classifiers Statistical-Computation	Exact Agreement Rate of 0.55	Style and Content
Schema Extract Analyse and Report	J. Christie	1999	Information Extraction Natural Language Processing	Correlation of 0.3	Style and Content
Intelligent Essay Marking System	Ming, Mikhailov and Kuan	2000	Pattern Indexing Neural Network	Correlation of 0.8	Content
C-Rater	Jill Burstein	2001	Natural Language Processing Statistical- Computation	Accuracy of 0.85	Content
MarkIt	Heinz Dreher and Robert Williams	2000	Multiple Linear Regression Vector space-computation	Correlation of 0.75 to 0.78	Content
Papers School Free Text Marking Engine	Mason and Grove-Stephenson	2002	Natural Language Processing Multiple Linear Regression	Not Given	Content
Bayesian Essay Test Scoring System	Rudner and Liang	2002	Bayesian Conditional Probability	Accuracy of 0.98	Style and Content
Automark	Mitchell et al.	2002	Information Extraction Natural Language Processing	Correlation of 0.93	Content
SAGrader	Idea Works	2010	Semantic Networks	Not Given	Content

Table 2.1: List of AEG systems

2.6 Comparisons between Human and Computer Markers

The main arguments for and against automated grading, gathered from several works, are listed below.

2.6.1 Advantages

When a human marker is given the task of grading a large number of essays, it is highly likely that the amount of attention given to each successive essay might possibly decrease during the course of marking. Hence, an automated system has a great advantage over manual human marking since it would not conceivably 'run out of energy'. Streeter et al. (2003) believe that the other advantages of an automated system include: a uniform objectivity since the computer is never subject to value biases; and its ability to analyse each essay with the same level of attentiveness without becoming bored, irritated or inattentive. Furthermore, lower attentiveness might prevent markers from recognising instances of plagiarism; whereas, a computerised system with access to a database of student answers can more easily detect this (Palmer, Williams and Dreher 2002).

Also, a marker is sometimes required to provide feedback to the student. This, coupled with the time that it takes to grade an essay, can tend to become an extremely tedious task and markers, after the first few batches might just neglect to provide constructive feedback, even though it could be important for the improvement of a student's writing. An AEG system would address this problem and, in addition, the materials from

the feedback could also be used to improve reading and communication skills (Godshalk, Swineford and Coffman 1996; Conlon 1986; Hearst 2000).

Finally, the reduced costs and overall improvement of the marking process is the common positive aspect of implementing an AEG system. Barring the initial installation costs, the system could be implemented across several departments, thereby reducing the long-term costs (Chung and O'Neil 1997).

2.6.2 Disadvantages

Researchers and markers alike have not hesitated to criticise the notion of a computer being given the task of grading student essays. Many believe that, given the myriad ways that any concept, story or point of view can be conveyed, it seems unlikely that a computerised system would be able to handle every possible written aspect of natural text.

Therefore, some critics opine that the system has limited capacity for accurate and valid assessment.. Moreover, Ford (2000) also stated that even with training data, a system might not be able to handle every type of question or answer.

Other criticisms of computerised grading include the opinion that there are still some things that only a human can do. More specifically, the computer lacks the common sense and intelligence possessed by a human marker, thus creating the impression that a score assigned by an automated system might not always be valid or credible.

2.7 Conclusion

The long-standing debate concerning the effectiveness of a computerised marking system as opposed to a human marker is well known. Researchers in the field have defended the use of an automated system, stating that these systems were not meant to replace human marking, but merely to facilitate it.

This chapter has described the different methods thus far that have been employed in the field of automated essay grading. Through a review of the literature, it has been found that the general trend of most systems developed for automated scoring favours the statistical-based, semi-supervised machine learning method.

It has also been found that statistical methods tend to have problems in dealing with tacit information. However, while systems that utilise artificial intelligence do perform better, the costs of running these systems are often quite high.

The next chapter formalises the issues and describes the objectives of this thesis. The research methodology chosen for this study is also described.

Chapter 3-Problem Definition

In the previous chapter, a review of the work done in the fields of Automated Essay Grading and Narratology, among others, was presented and discussed. It was established that while a significant amount of work has already been done in these areas, AEG systems and Narratology have yet to be combined.

This chapter will highlight the specific goals of this research and will also discuss the issues pertaining to Automated Essay Grading systems and also issues that need to be addressed before one can consider merging the concepts found in AEG and Narratology. Although several such issues have been addressed at some level by other researchers, this project attempts to tackle those issues using an innovative approach.

3.1 Introduction

While each essay grading system, when put through some sort of evaluation shows promising results with regards to the high correlations with human markers, it is here that there lies a fundamental problem: most of these systems already have an *a priori* result to go by, often a "model "or " best answer" type response. A human marker grades an essay according to his or her understanding and interpretation which is essentially subjective.

This is not to say that human markers are unable to carry out objective marking. This is probably the way all grading should be carried out; but to suppose that every other essay that is to be graded by a human marker would have to be first compared with the

'model' essay is just absurd; where then is the objectivity? How is it that when trying to automatically grade essays, it is seen as perfectly acceptable to grade them based on how closely each one approximates the model response?

From the perspective of a computer system, however, this of course makes perfect sense; why shouldn't an essay receive more or less the same grade if it has the same characteristics as another of the same grade? These approaches have all been empirically proven to work using the systems mentioned in Chapter 2 through the use of multiple linear regression and various other statistical and text processing measures that pick out features of an essay which have a high correlation to the grade that it receives. Having said that, when one takes a step back to view this process as a whole, one realizes that the process is in fact working backwards. While this is method would seem the most appropriate since it is an effective way to predict subsequent results, the effect of this is that, instead of looking at the content of the essay and determining its grade from there, the essay is graded based on a preconceived notion of what it *should* contain, not what it *does* contain.

3.2 Dealing with Tacit Information

Tacit information in the context of this thesis refers to contextual features within the text that are not recognised by a machine. Instances such as a certain character's reaction or emotional state are examples of such. This is not such a big problem in other types of systems such as those that deal with information retrieval or knowledge

acquisition, but contextual understanding is a rather critical area for automated essay grading.

Several systems in recent years have addressed the issue of contextual analysis (AutoMark, Intellimetric) with varying degrees of success. However, while analysing the content of a written report based on a certain topic is relatively easy, if the topic is already known, a lite-ontology can be constructed and measured against the essay's content, the same is not so for a narrative story. Scharf (2004) describes the modelling of a narrative domain, such as that of a story world, to be extremely difficult since there is an infinite number of possible settings and occurrences.

The main difference is that a story does not rely on the conventional method of discourse; a good story utilizes more of a showing than telling approach and allows the reader to seamlessly follow the course of events as they unfold in the story. This difference is also what makes the automatic grading of a narrative type essay so difficult, since narrative essays use more descriptive expressions that do not necessarily follow the more formal writing structures and language.

The immense scale of possibilities of descriptions that can be employed in free text makes it nigh impossible to model a logic framework that is able to comprehend the text. The sheer volume of data that is required to even undertake such a task would make it a highly impractical and economic nightmare. Having said this, it might still be feasible to create a general framework or a common sense framework that would allow a grading system to have a basic understanding of the text. This alone would

suffice to at least analyse the more implicit aspects of a narrative essay such as its structure, coherence, and the introduction and development of characters.

As mentioned earlier in Chapter 1, the scope of this thesis is limited to the more stylistic aspects of a narrative essay. The next section gives an overview of the NAPLAN marking rubric, followed by a description of the specific criteria this thesis attempts to tackle.

3.3 NAPLAN Marking Rubric

The scope of automated marking in this thesis will be limited to the Narrative Marking Guide of the 2010 National Assessment Program, Literacy and Numeracy (NAPLAN) rubric, which is detailed below:

The 2010 NAPLAN marking rubric is what is used to grade narrative essays of students in grades 3, 5, 7 and 9 in Australia. The rubric is made up of ten criteria, namely:

- Audience – writer's capacity to orient, engage and affect the reader
- Text Structure – organization of narrative features in an appropriate and effective structure
- Ideas – creation, selection and crafting of ideas
- Character & Setting – portrayal of character and/or development of a sense of place, time and atmosphere
- Vocabulary – the range and precision of language choices
- Cohesion – the control of multiple threads and relationships

- Paragraphing – segmenting of text into paragraphs that assist in reading
- Sentence Structure – production of grammatically correct, structurally sound and meaningful sentences
- Punctuation – use of correct and appropriate punctuation
- Spelling – accuracy of spelling and difficulty of words used

While each of these criteria is important in determining an appropriate score for a student's essay, it would be an immense task to apply all 10. Furthermore, it is possible to split these 10 criteria into two groups based on the different aspects of an essay they address respectively. Table 3.1 below shows a breakdown of these features and the specific criteria to which they relate.

Essay Aspects	**Rubric Category**
Stylistic	Audience Ideas Character & Setting Cohesion
Structure and Organization	Text Structure Vocabulary Paragraphing Sentence Structure Punctuation Spelling

Table 3.1: Breakdown of NAPLAN rubric according to Stylistic and Structure & Organisation

For the purposes of this thesis, the focus of the scoring model will be limited to the Stylistic aspects of an essay: specifically, the Audience, Ideas, Character & Setting and Cohesion criteria of the rubric, as it is for these aspects of the text that one would have to employ a more contextual approach to be able to determine a score. The other criteria related to Structure and Organisation is the subject of a complementary MPhil thesis.

3.4 Key Concepts and Definitions

The terms commonly used in the literature on narrative analysis include 'protagonists', 'antagonists', 'actions' and the like. Although there is a fuzzy distinction between some of these terms, it is imperative for purposes of formalisation to provide a clear distinction between them. In order to clearly define the concepts used in this thesis, this section will provide clarifications of the concepts used henceforth.

Most of the concepts that are detailed below are inspired by the idea of what constitutes an important, plot-driving sentence within a story. As such, the following concepts, unless otherwise stated, are considered in the context of a narrative essay/story.

3.4.1 Essays

Firstly, an essay can refer to any type of free-form text presented in a manner of the author's choosing. This can be in the form of an argumentative, persuasive, expository or narrative style. Other forms of essays may combine written text with graphic representations such as illustrations, photographs, diagrams or graphs which are

intended to enhance the point being made; other times it might just be a drawing that is meant to represent an idea but often times it is indecipherable by anyone else apart from the author.

For the purposes of this thesis, the scope will be confined to the narrative essay type. More specifically, this work will focus on the narrative type essays written by students in grades 1 to 12 for the purposes of examination by a human marker.

A narrative essay can be described as a story which contains a series of Events connected to one another, which are caused by or experienced by the characters of the story. Hence, an Event is an important part of the story since it furthers the plot.

In Kenneth Burke's Grammar of Motives (1969), he states that a good narrative should, as a minimum, contain an actor, action, goal, scene and instrument. In this work, it is taken that for a sentence to be considered important and thereby an Event, it should include at least an Actor, Action and State (scene).

3.4.2 Event

Definition – the encapsulation of an Actor, Action and a State. If any one of these is missing, a sentence is not classified as an Event.

Events in a narrative may depict plausible scenarios, such as experiences in a day in the life of a particular character, or they may go down the path of a fantasy in which the characters are magical creatures such as goblins or dragons with uncommon names and in unusual environments. Although there is a very clear distinction separating the

abovementioned genres, fundaments such as the inclusion of characters, remain the same.

3.4.3 Actors

Definition – a character, mentioned either by name or anaphora, as part of the story.

Previously, in the literature of narratology an Actor is said to be the character that performs an action, thereby being part of or the cause of an Event (Bal 1985). However, in this thesis, an Actor can be any character that is introduced to the audience or reader either by reference by another character or through introduction into the scene. This term is therefore not limited only to the character whose perspective is being portrayed, but also applies to other characters mentioned either by the author or another character. It is also important to note that an Actor need not necessarily be human.

3.4.4 Actions

Definition – an Action is an act that is performed by an Actor

Simply put, an Action, in this context, cannot occur without a cause. This cause is commonly, but not exclusively, represented by an Actor. In this case, an Action is always expressed with a verb.

3.4.5 State

Definition - the location, time or condition of the respective Actor

The State refers to the current situation, within the context of an Event which includes an Actor. This could be represented in a number of ways such as the Time, Location or Condition.

The Time is mentioned in terms of its passing or as a period in which an Event exists, or both. For example, “two weeks later” signifies the passage of time but also indicates the period when an Event takes place.

A Location refers to the physical location of an Actor or where the Event takes place. This is not limited to specific names defined in Named Entity Recognition tools such as countries or towns (proper nouns) but includes other locations such as “Jimmy’s house”.

A Condition refers to the physical or mental state of the Actor. Reference to the physical state of the Actor could depict an explicit injury such as a broken limb, or an implied injury such as the loss of blood or bleeding.

3.5 Marking Criteria

Each marking criterion that is included in the scoring model is split up into a specific number of bands which signify the quality of writing as determined by the rubric. This section will detail these individual bands as described in the NAPLAN rubric, thus providing more insight into the tasks involved in assigning a score based on these definitions.

3.5.1 Audience

Band	Description
0	Symbols or drawings which have the intention of conveying meaning
1	Contains some written content
2	Shows awareness of basic audience expectations through the use of simple narrative markers
3	An internally consistent story that attempts to support the reader by developing a shared understanding of context
4	Supports reader understanding and attempts to engage reader
5	Supports and engages reader through deliberate choice of language and use of narrative devices
6	Caters to the anticipated values and expectations of the reader Influences or affects the reader through precise and sustained choice of language and use of narrative devices

Table 1.2: Audience band scores under NAPLAN rubric

3.5.2 Ideas

Band	Description
0	No evidence or insufficient evidence
1	Ideas are very few and very simple
2	Ideas are few but not elaborated
3	Ideas show some development or elaboration All ideas relate coherently to a central storyline

4	Ideas are substantial and elaborated Ideas effectively contribute to a central storyline The story contains a suggestion of an underlying theme
5	Ideas are generated, selected and crafted to explore a recognisable theme Ideas are skilfully used in the service of the storyline

Table 3.3: Ideas band scores under NAPLAN rubric

3.5.3. Character & Setting

Band	Description
0	No evidence or insufficient evidence
1	Only names the characters or gives their roles (e.g. father, the teacher, my friend, dinosaur, we, Jim) and/or Only names the setting (e.g. school, the place we were at); setting is vague or confused
2	Suggestion of characterisation through brief descriptions or speech or feelings, but lacks substance or continuity and/or Suggestion of setting through very brief or superficial descriptions of place and/or time
3	Characterisation emerges through descriptions, action, speech or the attribution of thoughts and feelings to a character and/or Setting emerges through the description of place, time and atmosphere

4	Effective characterisation. Details are selected to create distinct characters and/or Maintains a sense of setting throughout. Details are selected to create a sense of place and atmosphere

Table 3.4: Character & Setting band scores under NAPLAN rubric

3.5.4 Cohesion

Band	Description
0	Symbols or drawings
1	Links are missing or incorrect Short script Often confusing for the reader
2	Some correct links between sentence (do not penalise for poor punctuation) Most referring words are accurate
3	Cohesive devices are used correctly to support reader understanding Accurate use of referring words Meaning is clear and text flows well in a sustained piece of writing
4	A range of cohesive devices is used correctly and deliberately to enhance reading An extended, highly cohesive piece of writing showing continuity of ideas and tightly linked sections of text

Table 3.5: Cohesion band scores under NAPLAN rubric

3.6 Research Issues

This section will define in detail the issues this research hopes to address based upon the review of the current literature.

3.6.1 Issue 1: Large amount of training data required

As shown in Chapter 2, most AEG systems require a large amount of training data. This is not a major issue when it comes to machine learning; however, one needs to consider that in most instances of grading, the questions do not always stay the same.

Generally, machine learning processes are designed to perform a specific task repeatedly using the same criteria after being trained on a set of data. In the instance of essay grading, however, it cannot be assumed that the domain in question remains constant. Therefore, unless the essay is marked solely on its technical features, its grade cannot be calculated unless the domain concept of the marking system is altered and trained again. This could be a serious problem should the system be based on a logic process that is manually tagged and trained.

This issue is more significant when a small number of essays is being graded. Usually, AEG systems are used to grade a large number of essays which significantly reduces the amount of time and resources expended.

The amount of training data that is required in that instance would therefore be only a small fraction of the total (currently, the most training data that is required by an AEG system is the E-Rater at a minimum of 270), but the economic benefits of such a system are significantly less when there are fewer than 100-200 essays to be marked. Hence, a

system that is able to eliminate or at the very least reduce the impact of this restriction would be a huge step in advancing this technology.

The problem of requiring a large amount of data also brings to light the problem of acquiring a large body of essays to work with in the first place. Furthermore, according to experts there is also difficulty involved in obtaining a corpus of essays where all grades are agreed upon (Valenti 2003; Larkey 2003). Moreover, student essays are often handwritten; the transferring of data from hard to soft formats requires either manual labour or advanced software which are, respectively, time consuming and expensive.

3.6.2 Issue 2: Errors in scoring

Systems that formulate the scoring algorithm based on the appearance of certain key words or the like are at risk of being 'fooled' by anyone who has even a basic knowledge of how they work. This 'counting' method is similar to the Bag-of-Words and Named Entity Recognition approach whereby a high occurrence of certain key words contributes to a higher score, even if those words are not used within the correct context of the subject domain. Hence, a 'nonsense' passage that contains many words that are related to the subject domain, despite being a rather naïve approach to fool most well-designed systems, still carries the possibility of influencing the grade (Ford 2000).

Reliance on NER also has its inherent problems when it comes to a narrative essay. Apart from the entity boundary problem wherein an entity can be represented by more

than one word, locations in a narrative are usually not as simple as “London” or “Townsville”; instead, more informal descriptions such as “a dark cave” or “a sharp cliff” are more common, making it difficult for an NER tool to pick out.

This is refuted by the argument that if a student is able to fool the system into awarding a good grade, that student must already have a good grasp of the domain and thus deserves a good score anyway (Dessus et al. 2000). Albeit a logical point, the onus is on the system to be able to detect such occurrences and flag them out should they appear. Besides, word counting measures, while adopted by some systems, are not always a good indication of technical features such as grammar and sentence structure.

3.6.3 Issue 3: Incomplete scoring methods

In cases where the system places more emphasis on the technical characteristics of the text, other aspects of essay writing get overlooked. As mentioned earlier, most systems adopt a feature extraction method which is used to check for similarities with model answers to predict an essay’s score. This method totally ignores the more tacit traits of a student with a good mind for discourse in written language. On a simpler note, this would not allow the system to evaluate how well a story flows or how well a certain scene is presented to the reader.

Palincsar et al. (1994) reiterate the fact that the computer is incapable of many things that a human can do, such as understanding wit or sarcasm that are often a more subtle approach to engaging the reader.

This problem has not been addressed by current literature and will probably not be for quite some time since formalising the tacit aspects of writing such as sarcasm can be too difficult, considering that even some human markers would have trouble identifying it. The best that has been done so far is the creation of a conceptualised framework of a knowledge domain which places text into context by means of certain parsing tools, but this constricts the grading system to one specific domain at a time.

3.6.4 Issue 4: Structured input is taken for granted

Common methods of text analysis or extraction methods take in specific forms of input such as news articles or reports and journals, which are mostly grammatically and structurally correct samples of unstructured text. For the task of information extraction and other forms of analysis (such as subject-domain), this would be ideal since the system would not have to deal with anomalous data with large amounts of spelling errors or the like.

However, it is not practical to assume that every essay to be marked is free of spelling and grammatical errors. Furthermore, although a large number of such mistakes would affect the overall grade given, it does not necessarily mean that the essay content itself is not coherent or engaging.

Another problem with regularly-occurring mistakes is that the system might not be able to efficiently parse the textual content of an essay should a keyword be misspelt. A human marker would be able to pick out a spelling mistake and still be able to grasp

the intent of the author; this is one of the major hurdles to be overcome in developing an essay grading system that can function as well as a human grader.

3.6.5 Issue 5: Highly dependent on domain

Most of the AEG systems that try to deal with the context and content of an essay would need to have an extensive knowledge representation of the subject domain; possibly an ontological representation. This would in turn produce the problem of AEG systems being too highly focused on a single domain. This is a significant issue when trying to analyse a narrative essay since, even if the students are given a specific topic to write on, it is not usually one wherein the subject domain can be easily modelled.

In order for an AEG system to conduct an analysis of the more tacit aspects of natural text, there needs to be a conceptualisation that can understand it regardless of the subject domain - in other words, a high level analysis that allows a more specific examination of an essay's content.

3.7 Research Aims

Having detailed the issues found through a review of current literature in the section above, the next section will discuss the aims of this project.

3.7.1 Aim 1: Creating a Semi- Domain Independent Model

The problem of requiring a large amount of training data, together with the problem of having to train a system repeatedly on different subject domains has been an issue that for many years has plagued the field of automated assessment.

A very obvious solution to this is to create a system that is able to grade an essay regardless of the subject domain. This might seem like a very naïve and overly-ambitious approach to the solution, but when this is applied to a narrative type essay, it actually becomes more feasible. There is no practical way to predict which subject domain a student will choose when writing a narrative, even with a given title and subject prompt; therefore, it would be rather useless to train the system on a specific subject domain.

Therefore, one of the main aims here is to design a system that is able to assign a fair grade regardless of the subject domain. This would be hugely beneficial since it would not require a large amount of training data while at the same time reducing the overall process time of the system itself.

3.7.2 Aim 2: Picking out Tacit Features

The problem here is that, when carrying out contextual analysis of natural text be it through noun/verb phrases or text parsing, in order to understand the contextual meaning of a word, all senses of the word should be known beforehand so as to put that word into context. The creation of such a lexicon would be extremely resource-intensive, not to mention the enormous amount of maintenance required in sustaining it.

In order to pick out the more subtle aspects of the text without having to create a massive knowledge framework, the system should be able to pick out instances within

the text that relate to those subtle aspects. Therefore, the aim here is to design a methodology whereby this can be done in a feasible manner.

3.7.3 Aim 3: Creating an In-depth Scoring Model

The above aim would then directly lead to the formulation of an in-depth scoring model. Although this has been somewhat achieved using Latent Semantic Analysis (Landauer and Foltz 1998) and Artificial Intelligence (Rudner et al. 2006), those solutions involve heavy computations that consume a large amount of resources that are not readily available.

In creating a feasible model of text analysis that picks out stylistic features, coupled with the evaluation of technical features that have already been done, an in-depth scoring model can be designed without placing too much of a burden on available resources.

As mentioned earlier, the focus of this thesis is on four criteria, namely:

3.7.3.1 Audience

This criterion determines the extent to which the narrative essay is able to immerse the reader in world of the story. As such, it is important that the story has an uninterrupted flow with no or minimal gaps in between. The reader should also be able to follow the story easily; therefore, events should not be haphazardly strewn about the story, but instead should be ordered in such a way that they engage the reader.

Therefore, the aim here is to pick out those features of the text that engage a reader.

3.7.3.2 Ideas

In this criterion, the attention is applied to the creation and crafting of ideas for a narrative, or in other words, how an Event is created and applied within the story. Assuming that an essay can be broken down into individual sentences, in order for a sentence to be considered an Event, it has to include three concepts: Actor, Action and State. Hence, in order to score against this particular criterion, it is necessary to identify those sentences which contain these concepts that constitute an Event.

3.7.3.3 Character and Setting

It should be noted here that in the context of the NAPLAN marking rubric, this criterion is actually meant to indicate the presence of a Character, which is the portrayal and development of a character in the narrative *and/or* the Setting, which refers to a sense of place, time and atmosphere. It is therefore not necessary for a narrative essay to contain both.

The proposed solution should have a method of detecting both occurrences and be able to assign a score accordingly.

3.7.3.4 Cohesion

The full description of this criterion is as follows: “the control of multiple threads and relationships over the whole text, achieved through the use of referring words, substitutions, word associations and text connectives” (NAPLAN 2010 pg. 6).

Therefore, in terms of this description, the scoring framework of the proposed solution should be able to not only identify these “connectives” and other referring words, but also their appropriate use.

3.8 Summary of Problem Definition

Recapping the main points of this section, the purpose of this thesis is to create a scoring model that is able to:

1. Design a system that is primarily based on a set marking rubric. In this case, the NAPLAN marking rubric, specifically evaluating an essay’s score using the stylistics aspects of the rubric which were:
    a. Audience
    b. Ideas
    c. Character and Setting and;
    d. Cohesion.
2. Create a framework for automated scoring that does not rely heavily on the presence of training data.
3. Ensure that the scoring framework is not reliant on the subject domain.

Although there are still further issues that might need to be addressed in current AEG systems, the ones mentioned in this chapter form the foundation upon which the proposed model will be built. Here, the aim is to be able to pick out the more subtle layers of written text by focusing on its stylistic aspects.

Since it is these layers that showcase the author's understanding of ways to express knowledge, or immerse the reader in the story world, the ability of a grading system to be able to consider them whilst assigning a grade would be a tremendous step towards a more robust scoring system.

The research methodology to be undertaken for this research venture will be discussed in the next section.

3.9 Research Methodology

The aim of this section is to provide a brief overview of the chosen methodology on which this research project will be based. In carrying out research that can be considered valid in its respective discipline, it is important that it be based on a sound research methodology. To achieve this goal, this section discusses a number of available research methodologies and justifies the choice of the particular methodology chosen as the most suitable for the purposes of this research and its desired outcomes.

3.9.1 Research Approaches

Research methods can be grouped under two categories, namely:

- Social Science; and
- Science and Engineering

3.9.1.1 Social Science

The social science approach mainly involves, as the name suggests, the social aspects of the research such as ideas and concepts. The stages of this research method usually

involve “action, role-playing and descriptive research and reviews” (Galliers 1991). This approach can be further broken down into two sub-categories:

- Quantitative; and
- Qualitative

The first, Quantitative, is commonly applied to research that follows the process of first having an initial hypothesis, generally in relation to the existence of relations or correlations between certain measurable variables. Here, the researcher assumes that there are a number of different interpretations or viewpoints pertaining to the particular subject at hand. Thus, the main goal is to determine if there are in fact any measurable relationships between these variables and, if there are, a method with which to detect and measure them. As stated by Juristo and Moreno (2002), the goal of this methodology is to determine if there exists a numerical relationship between said variables.

To determine this, one usually has to gather a large data sample by various means such as interviews or questionnaires. Several statistical analysis methods are then applied to the data with the goal of either proving or disproving the given hypothesis.

Conversely, Qualitative research is more concerned with the ‘how’ and ‘why’. As opposed to the ‘yes’ or ‘no’ assumption of the former, this methodology is based on the assumption that there may be many paths to the solution to a problem, rather than one. Furthermore, it may also be true that each of these solutions is equally valid or true (Creswell 1998; Guba and Lincoln 1989).

3.9.1.2 Science and Engineering

On the other hand, the science and engineering approach is based on gathering empirical and measurable evidence through observation and experimentation, together with the formulation and testing of certain hypotheses.

Scientific research methods include laboratory experiments, field experiments, surveys, case studies, theorem proof, forecasting and simulation and are usually distinguishable by their "repeatability, reductionism and refutability" and assume that "observations of the phenomena under investigation" can and should be made objectively (Galliers 1991).

Research in the science and engineering field is said to tackle what is regarded as 'wicked' problems (Rittel and Webber 1984). Problems classified as such usually have the following characteristics, adopted from Hevner et al. (2003, p. 10):

- "Unstable requirements and constraints based upon ill-defined environmental contexts"
- "Complex interactions among subcomponents of the problem and its solution"
- "Inherent flexibility to change design processes as well as design artifacts (i.e. malleable processes and artifacts)"
- "A critical dependence upon human cognitive abilities (e.g. creativity) to produce effective solutions"
- "A critical dependence upon human social abilities (e.g. teamwork) to produce effective solutions."

This method is essentially a problem solving process (Hevner et al. 2003). This form of research typically includes the application of algorithms, human/computer interfaces, design methodologies (including process models) and languages. Its application is most common in the field of Engineering and Computer Science, although it can be found in many other disciplines and domains (Vaishnavi and Kuechler 2005).

Inherently, the science and engineering based research approach can be split into three levels (Nunamaker et al. 1991; Galliers 1992; Burstein & Gregor 1999):

- Conceptual level - creating new ideas and concepts through analysis and design processes.
- Perceptual level - formulating a new method and approach by designing and building the tools or environment or system through implementation. This stage forms the conceptual framework, which is the foundation of the end product and should be constantly referred to when working on other stages. This is the more important aspect of the process as it is in this stage that most of the primary concepts are developed. As stated above, this stage should be constantly referred to even while progressing through the other stages.
- Practical level - carrying out testing and validation through experimentation with real-world examples, using laboratory or field testing. Evaluation and validation of the end product gives valuable feedback information on its effectiveness and accuracy, enabling researchers to improve on the overall process in addition to enhancing the quality of the end product.

3.9.2 Choice of Research Methodology

The previous section listed several ways in which a research problem can be tackled. The Science and Engineering methodology is based on designing a solution to a given problem or problem areas. The quantitative and qualitative aspects of Social Science methodologies on the other primarily deal with proposing a hypothesis and trying to determine if the data collected proves or disproves it, or to understand the more subtle concepts such as the 'why' and 'how', respectively.

There are some instances, however, where multiple methodologies are needed to solve a problem. This is one such instance. Since this thesis primarily deals with the development of a computerised system, the Science and Engineering methodology will be used as the primary method of choice.

However, with the inclusion of tacit features of text that require a more subtle process, the qualitative aspect of Social Science research methodologies will also come into play. Hence, this thesis will adopt a hybrid methodology which will be applied to the scope of the issues to be addressed.

3.10 Conclusion

This chapter discussed the issues that were identified through a review of the literature in the field of Automated Essay Grading. The main goals of this thesis were described step-by-step and led to the choice of the appropriate methodologies to be used to tackle the problem areas identified. The key concepts and definitions that will be used throughout this thesis were also listed and discussed.

Thus far, to the best of our knowledge, no grading systems have been developed that are based primarily on specified marking rubric. It is hoped that in doing so, a new approach to Automated Essay Grading may be discovered that, despite its being limited for the time being to the narrative essay type, will be able to grade essays independent of their subject domain.

The next chapter will present the overview of the conceptual framework of the proposed solution.

Chapter 4-Theoretical Framework

This chapter presents the theoretical framework and an overview of the solution. It begins with a detailed view of the concepts of Narrative Analysis that eventually led to the formulation of the proposed solution. This is followed by an overview of the method by which the proposed solution will address the problems described in Chapter 3, together with a description of the tools used in performing the textual analysis processes.

Prior to this, Chapter 2 examined previous work that had been done in the context of Automated Essay Grading, which revealed that AEG systems would benefit from a more comprehensive scoring model that incorporates certain concepts of Narratology, which has not been attempted previously.

4.1 Introduction

As mentioned earlier, the scope of the scoring model encompasses the four criteria that are most suited to contextual analysis. In order for this to be realised, there needs to be a clear understanding of the steps involved.

Chapter 3 identified the following three main research aims:

For Automated Grading System purposes:

4. Create a framework for automated scoring that does not rely heavily on the presence of training data.
5. Ensure that the scoring framework is not heavily reliant on the subject domain.
6. Evaluate an essay's score using the framework according to the criteria of Audience, Ideas, Character & Setting and Cohesion as specified by the NAPLAN Rubric.

One of the main problems of some Automated Marking Systems is the need for copious amounts of training data. In addition, should there be a new subject domain introduced into the marking scheme, the system would again have to be trained on a certain number of essays before being able to assign an appropriate score to an unmarked essay.

Therefore, in order to create a scoring system that does not have to be constantly retrained, a viable solution is to create a framework that needs to be trained only once. This leads us to the second aim, which is domain independence.

In dealing with the second aim, to ensure that the framework is not heavily dependent on a particular subject domain, it is necessary to focus most of the analysis on the layer of the text concerned mostly with how the text is presented to the reader. Thus, there is less need to delve into the subject, negating the need for a heavy dependence on any specific subject domain.

In addressing the method by which an essay receives a score, as mentioned in Chapter 3, each criterion has specific descriptions as stated by the marking rubric. Hence, the

chosen solution needs to be able to interpret how those descriptions would be related to the essay in terms of identifiable features within the text.

In terms of textual analysis, the intention is to create a system that is able to:

- perform an analysis of textual content while being able to draw some contextual inference;
- execute on any machine without using a copious amount of resources; and,
- be run without first having access to 'pre-knowledge'.

Before the proposed solution is presented, a detailed discussion of the concepts of Narratology that have influenced the formulation of the solution is provided. Towards this goal, the next section of this chapter gives a more detailed description of the field of Narrative Analysis. It is hoped that by studying the concepts therein, a novel method can be developed that incorporates these concepts in a system of automated essay grading.

4.2 Narrative Analysis

Since the scope of this research project includes the analysis of a narrative type text and its elements, a brief review of work done in the field of narratology would give more insight into the formulation of a possible solution. The sections below will first describe the basic concepts of a narrative text, followed by a review of current work done in the field of narratology.

4.2.1 Narratives

What is a narrative? There are dozens of definitions of what a narrative is, what it should encompass and what it should do. Many of us might scoff at that very question since it seems obvious that a narrative is merely a block of text, presented to a reader that recounts certain happenings. In fact, traditionally a narrative text can be said to contain a series of events, whether fictional or not, recounted from the perspective of the author or one or more main characters, and told in the first or third person; narrator to narratee. Indeed, the Merriam-Webster's online dictionary defines a narrative as "the representation in art of an event or story", which might lead one to conclude that the narrator or narratee might be the glue that holds it all together through a recount of past events.

What are the features of a narrative? In this work, it is believed that the core of a narrative is the 'event'. As succinctly defined by Abbot (2002), a narrative is "the representation of an event or a series of events". He goes on to say that without an event, a block of text may be a description, an exposition or even an argument, but never a narrative. This will be further discussed in the later sections but first, in order to better understand the intricacies of a narrative, it is necessary to look at its different layers.

4.2.1.1 Fabula and Sjuzet

Mieke Bal, in his work on narratology mentions the fabula and sjuzet, terms coined by Vladimir Propp and Shklovksy to describe the constructs of a narrative (Bal 1985). In many works involving narratology, these two terms may be mentioned several times,

although they are often represented in several ways. Bal's description of the fabula involves the logical and chronological order of events as they are presented to the audience. Although one would commonly picture a film or book playing this role through discourse, it could in fact also be done without it, much like a photo montage or a painting would. This is commonly confused with the next layer, which is the sjuzet.

The sjuzet, a term preferred by many scholars when discussing narratology, is the particular way in which the events are presented, more commonly referred to as the 'plot' or 'story'. This might seem confusing at first since when one is talking about the fabula; it would inadvertently seem that they were talking about the story. But remembering what was mentioned in the above paragraph, the sjuzet differs from the fabula in that the former is the *way* the story is told, regardless of the chronological order of the events. For example, in a film that tells the story of a person's life, the director might show certain flashbacks or flash-forwards to present an event to the audience. As such, when an event is shown as a flashback, the audience would know that the event was in the past and when the scene is over, the story is back to the present. Therefore, the fabula is the way events are ordered in the story, but the sjuzet in contrast, is the way the story is told.

The third and topmost layer is the narrative itself. The narrative is its structure or form, derived from all the parts that make up the narrative. As such, the topmost layer could take the form of a textual, pictorial or multimedia representation. Figure 4.1 below illustrates the levels of narratology.

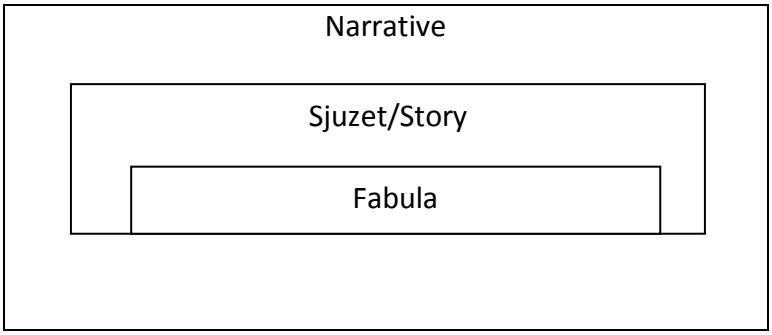

Figure 4.1: Layers in Narratology

In terms of these concepts, Mike Bal (1997) describes three distinct characteristics of a narrative text:

- There are two main types of spokesperson, one which is within the fabula, and one which is not
- There are three distinguishable layers, namely the text, the story and the fabula
- The text can be represented as a series of events caused by or experienced by the actors

It should be noted that works based on the structuring of a narrative are also influenced by the different layers that are within. For example, automated grading systems that base their analysis mostly on the way events are ordered can be said to be concerned with the fabula. On the other hand, should the system's analysis be based on how those events are structured, the system would then be more focused on the sjuzet, or story. Lastly, systems that deal with the effect of the text or film or whichever form the final presentation takes are said to deal with the topmost layer itself, the narrative.

Understanding these layers is important in allowing one to obtain a general overview. In the later chapters these will be further broken down into more specific terms but for now, the works of other researchers on the subject of narratives will be described. Over the past decade, several methods have been used to analyse the narrative text type, such as via ontology and through the analysis of certain aspects of events throughout the text, all of which are described in the following sections.

4.2.2 Ontology-based Analysis

The term 'ontology' was initially used in the domain of philosophy to describe the study of the nature of being or existing. However, in the field of information technology, an ontology according to Gruber (1993) in his most quoted explanation is "an explicit specification of a conceptualization".

Maedche and Staab (2009) state that ontologies serve as a form of "metadata schemas, providing a controlled vocabulary of concepts, each with explicitly defined and machine-processable semantics". Generally, the benefit of this is that one can create a knowledge base that can be merged with another expanding its domain and scope; it is also possible to pick out certain parts from different ontologies to create a more refined knowledge base.

One might argue that since there already exists a multitude of thesauri and lexicon that pre-date ontologies and seem to serve the same purposes by providing a controlled vocabulary, why is there a need for the development of an ontology? Two reasons given by Witte et al. (2007) are that they give representational capabilities to

information retrieval and extraction tools; that is, they are able to identify relations between concepts by considering the actual context in which each concept is used (a richer and more informative representation of the concept); secondly, it improves semantic consistency by enabling better portability when trying to integrate other ontologies.

4.2.2.1 Narrative Ontology

In the field of narratology, Nancy Green proposed developing a narrative ontology for use in Artificial Intelligence in the domain of the narrative arts (Green 2002). In her work, she describes building the ontology using the characterisation of different goals, methods, symbol systems, participants in and results of what are called 'artistic creations' as the foundation for the concepts that will be used in the resultant ontology. Among the many concepts proposed, causal-temporal chain of events and story characters within the sub-heading of story worlds were included.

Henrik Scharfe (2004) also discusses a narrative ontology modelled on the concepts found in a narrative. The objective of Scharfe's work is to create a model that takes into account the structure and methods of interaction in the narrative domain, which according to him is both highly specialised and general at the same time. This is the basis for a framework in which an ontology of narratological terms may be organised (Scharfe 2004).

In creating his model, Scharfe tries to address three criteria: Firstly, it has to be general enough to incorporate a substantial part of the underlying theories seen in a narrative

type text. Secondly, while trying to remain general, the model should at the same time be specialised enough to account for characteristic features which are unique to the narrative domain. Finally, in order for there to an understanding of the concepts themselves, there must be stable categories that allow the correct categorisation and organisation of those concepts. In doing so, the chosen model framework was adapted from Jahn (2003), which consists of three levels of communication namely:

- Author to reader
- Narrator to narrate
- Character to character

The resultant ontology consisted of 631 concepts gathered from surveying the previous works of others such as Gerald Prince in 1987, Martin McQuillan in 2000 and H. Porter Abbot in 2002. The resulting ontology covered a large number of key narratological concepts, including their basic notions and derived forms. Scharfe believes that if used in conjunction with other narrative knowledge frameworks, a narrative ontology could form the basis for representing a large body of concepts consistently and coherently.

It was mentioned earlier in this section that narrative analysis can be done according to the three different layers identified by MiekeBal (1987). Tuffield et al. (2006) in their work outlining a simple taxonomy of the different approaches to narrative modelling, describe the following approaches from the lowest to the highest levels.

4.2.2.2 Models of the Fabula

Annotations of multimedia items, made possible by semantic web technology that constitute a body of knowledge, can be regarded as a form of fabula modelling, since the annotations contain information about the events and actors therein. Tuffield et al. (2006) give an example of annotating a video with information regarding the sequence of events as they happen in chronological order. In this instance, the description of the events would be independent of the story (how they appeared and their timeline within the story). Annotations of these details would allow one to construct a narrative through some reassembly of their presentation, which is similar to the work of Bocconi (2005), in which the author generated video documentaries based on rhetorical annotations.

Apart from this model of the fabula forming the basis for generating new stories and analysing current ones, the model could also be useful in presenting the raw information about the narrative (Tuffield et al. 2006).

4.2.2.3 Models of the Sjuzet

An ontological model of the sjuzet or story layer would have to focus mainly on the structure and arrangement of the events. Shneiderman (1997) mentioned that readers would have certain expectations on how the sjuzet should be, most often based on the genre. For example, for a typical “heroic quest” theme of the fantasy genre, one would usually expect the story to follow the general trend of first introducing the main character, followed by the object of the quest and eventually the ultimate conclusion of the quest. Tuffield et al. (2006) state that it is this structural knowledge that needs to

be modelled, wherein story threads play a crucial role. The quality of the annotations regarding this structural knowledge is also crucial, since being able to identify relationships within the knowledge base is largely dependent on them.

It has been proposed by some that in order to fulfil reader/audience expectations, when creating a model of a higher level knowledge base, it is necessary to use story grammars implemented through templates (Alani et al. 2003). These templates would allow one to work around the usual restrictions about how a story should be arranged according to its specific genre by defining a structure which is populated by the contents of the fabula. Although this allows one to bypass one restriction, it has drawbacks.

Using a template often means forgoing more flexible input methods. Therefore, static templates would have to be predefined by the system developer before the system itself can be deployed, thus constricting the model in terms of identifying new relationships and adapting to the content of the fabula (Tuffield et al. 2006).

4.2.2.4 Models of the Narrative

The narrative itself might seem a little too obvious to need to be involved in ontological modelling but, nevertheless, even with the fabula and sjuzet it is still necessary for the text itself to be presented through some form of medium, which is then perceived by the readers or audience. Tuffield et al. (2006) state that there are still semantic effects when faced with presentation choices, such as in cinematography where different methods of presentation such as a slow fade would be used to indicate the passage of

time. It would thus appear that an ontology modelled on this layer would be highly connected to and dependent on the form of the text.

4.2.3 Temporal Order Analysis

Sometimes it is necessary, apart from identifying within a narrative text which sentences are important events and which ones are not, to be able to order them temporally. The chronological ordering of narrative events is the more common approach since this type of ordering is defined by the appearance of said events as they are presented. However, the ordering of those events in a temporal fashion requires contextual understanding of the text so as to determine the correct order of the occurrence of events.

Nouioia (2008) attempts this temporal ordering with regards to texts describing road accidents. Here, the author attempted to automatically temporally order the events using a variety of methods which involved the use of temporal references that shared a link between specific words in the text. After obtaining an initial order or events through temporal information gathered from a semantic representation of the text, the default ordering is further modified by an algorithm that considers precedence and simultaneity constraints. The output gathered using this approach takes a linear temporal format that was adequate for determining which event happened before which, thereby establishing the cause of an accident.

Taking an alternative approach to narrative analysis, Chambers and Jurafsky (2008) believe that narratives are centred on a main character or protagonist, this being the

focus of a narrative chain. The authors attempt to identify this chain, defining a narrative chain as “a partially ordered set of events that share a common actor”, wherein a narrative event is represented by the actors involved, stated by the authors as a typed dependency. In evaluating the model of event detection, the authors borrow the concept of the cloze task, in which they name the narrative cloze. The narrative cloze is where in a sequence of events, one is removed and the task is to determine which one.

In conducting this method of analysis, the authors address narrative chain recognition through a series of steps:

- Narrative event induction - This step involves using an entity based model for learning narrative relations focused on the protagonist. Each event within the narrative serves to characterise the role played by the protagonist, with resulting relations to connected events. This is done using an unsupervised approach based on co-reference as evidence for the relationship between events.
- Temporal ordering of events - Having identified the events it is still necessary to order them into a narrative chain. The authors attempt to conduct a partial temporal ordering of events using a two-stage approach:
    - In the first stage, the model uses a supervised machine learning approach to annotate the attributes of events based on their temporal aspects, garnered from tense, grammar and other language

conventions. The classification of events' temporality is based on POS tags, neighbouring auxiliaries and WordNet synsets.

- The second stage involves further classifying the relationships identified using the output from the first stage combined with other linguistic features.

The task of detecting events in a narrative could also be applied to other fields of information extraction, such as text summarization. In the work of Enokiza et al. (2008), the authors look for events in important sentences for the task of text summarization. Toward this goal, the way that humans would comprehend a narrative was taken into account; what a human reader would deem as an important sentence in the text was taken to be the baseline for extraction. The results of two experiments showed that it was possible to detect several patterns that those sentences share through a series of connected propositions. The second experiment was conducted to see if those rules could be applied to accurately detect new events in an unannotated narrative.

4.2.4 Causal Relation Analysis

Within a narrative, Events are related to one another via Causal Relations. The Merriam-Webster's online dictionary defines causal relations as "the relation between a cause and its effect or between regularly correlated events". Among the first to take this into consideration was Trabasso and Sperry (1985), in which they introduce the concept of causality, where the reader would have to apply real-world knowledge to derive causal inferences, which tells the reader which event caused which. Although

this might seem similar to other works that focus on the temporal order of events, searching for causal relations take a less linear approach. For the temporal ordering of events to be identified, it has to be assumed that a specific event must take place within the scope of temporality, before another; moreover, an event might share a causal relationship with more than one event. For example, consider the following sentences:

(A) Kirk was admitted to the hospital

(B) where the doctors strapped him to the bed.

(C)Not liking to be restrained,

(D)Kirk struggled so fiercely,

(E) he broke his arm

Following the examples given by Trabasso and Sperry (1985), we can identify six main causal relations:

- Motivation Cause
    - Goal-orientated actions
    - Relation must be between goal and an action

      E.g. C is a motivational cause for D. In this case, not being restrained is the desired state
- Psychological Cause
    - Non-goal orientated actions

E.g. B is a psychological cause for D, since the original act of strapping Kirk has the psychological effect of Kirk not liking the present state.

- Physical Cause
    - Involves naïve interpretations of the physical world or of mechanical causality between objects and people.

      e.g. D is a physical cause for E

- Enablement
    - Actions, occurrences or states which are necessary but insufficient to cause other actions or states

      E.g. the action of admitting Kirk to the hospital (A) ***allowed*** for the following actions to occur but did not cause them to occur.

- Temporal Succession
    - Where in two events/actions occur successively but are not the cause of one another
    - Common is descriptions of characters, locale and setting

- Temporal Coexistence
    - Wherein two events/actions occur at the same time within the story but are not the cause of one another

- Also common in descriptions of characters, locale and setting

Girju and Moldovan (2002) proposed that a more direct approach is to look for cause (noun) and effect (verb) patterns as the first step in identifying a causal relationship. For example, the cause and effect pattern is explicit in the sentence:

"The earthquake caused the building to collapse"

Similarly, Chang and Choi (2004) mention in their works a 'Cue Phrase', which denotes a phrase or a group of words which through some sort of relationship, connect Events with one another. In the sentence mentioned above, the verb "caused" indicates this relationship. Girju and Moldovan (2002) also describe three other causal relations patterns that frequently occur:

- Explicit Effect noun and Implicit Cause noun
- Explicit Cause noun and Implicit Effect noun
- Both Cause and Effect nouns are implicit

One of the first works to identify verbs based on causation was that of Nedjakov and Silnickji (1973), in which they distinguish three categories of causative verbs:

- Simple causatives – wherein the linking verb is explicitly a cause. E.g. Dams generate electrical power; the verb "generate" being synonymous with cause.
- Resultative causatives – a linking verb that includes or is a part of the resulting situation. E.g. to kill, to cause death

- Instrumental causatives – a linking verb that is part of an instance as well as the result. E.g. to poison (killing by poisoning), resulting in death

The term 'narrative text' seems to have a rather generalised definition when it comes to works concerning narrative data as an input, ranging from the common idea of novels and story boards to more unconventional blocks of text such as news reports. While it is difficult to find any similarities when it comes to the specific focus of analysis, there is one characteristic they share in common: a narrative is mostly seen as a collection of events.

In other words, it is necessary for 'something to happen' for there to be a narrative. This is sometimes the main focus of works that aim to perform some sort of summarization or event extraction on narrative type texts. This work will also follow largely the same concept of analysing narratives, which is to focus more on the events or 'happenings' within the narrative.

The next section will discuss the proposed solution in which the issues identified in Chapter 3 will be addressed.

4.3 Proposed Solution

The chosen solution to achieve the aims listed at the beginning of this chapter is the Event Detection framework. Since a narrative text is essentially a series of connected Events caused or experienced by the characters, a method of formalising these Events would allow the text to be more easily interpreted by a machine. This method of

analysis will also focus largely on the fabula layer, which can be said to be the first layer of a narrative text. This would mean that there would less need to delve into the subject of the text, thereby allowing the system to remain independent of any subject domain.

For this purpose of Event detection, we take the criteria mentioned by Burke (1969) that make up an Event, where he states that a narrative should at a minimum contain an actor, action, scene and instrument. In the work in this thesis, it is assumed that the scene and instrument in this instance can be combined to form a current 'state' of things with relation to the Actor or Actors involved in that specific Event.

The Event Detection framework is intended to select specific features of a narrative text and, regardless of the subject domain, translate that information into an appropriate band score according to the marking rubric. In other words, it is a method by which the implicit features of the text are converted into explicit, machine-readable ones.

4.3.1 Overview of Proposed Solution

To briefly recap, it has been established that the proposed solution is centred on the detection of Events within an essay. However, simply detecting these Events is not enough to generate a grade which relates to each of the four criteria that are the concern of this thesis.

Before we can correctly grade an essay according to the aforementioned criteria, we must determine how the Events are related to the criteria themselves. This is done in

the Score Grouping stage of the solution, which identifies which features of the essay are most relevant towards the grade an essay receives under the respective criteria of Audience, Ideas, Character and Setting and lastly Cohesion. In so doing, we are then able to map certain features of the essay, so that the criteria pertaining to Events are used in grading the essay.

However, raw text is hardly suitable for machine processing; therefore, before any of the above analysis stages can be carried out, the raw text needs to be converted into a machine-readable output. This is done in the Text Analysis Stage, where Natural Language Processing tools are applied to the raw text to make it machine-readable. This output is then used in the Score Grouping stage which consists of the Event Detection and Rubric formalisation processes.

The Event Detection takes in all the output from the Text Analysis stage and uses the gathered information to determine whether or not a sentence can be classified as an Event. The Rubric Formalisation process mainly performs what was mentioned earlier, a mapping of the various features within the essay that relate to the grade it receives according to the respective criteria of Audience, Ideas, Character and Setting and Cohesion.

Figure 4.2 shows the overall view of the proposed solution's theoretical framework.

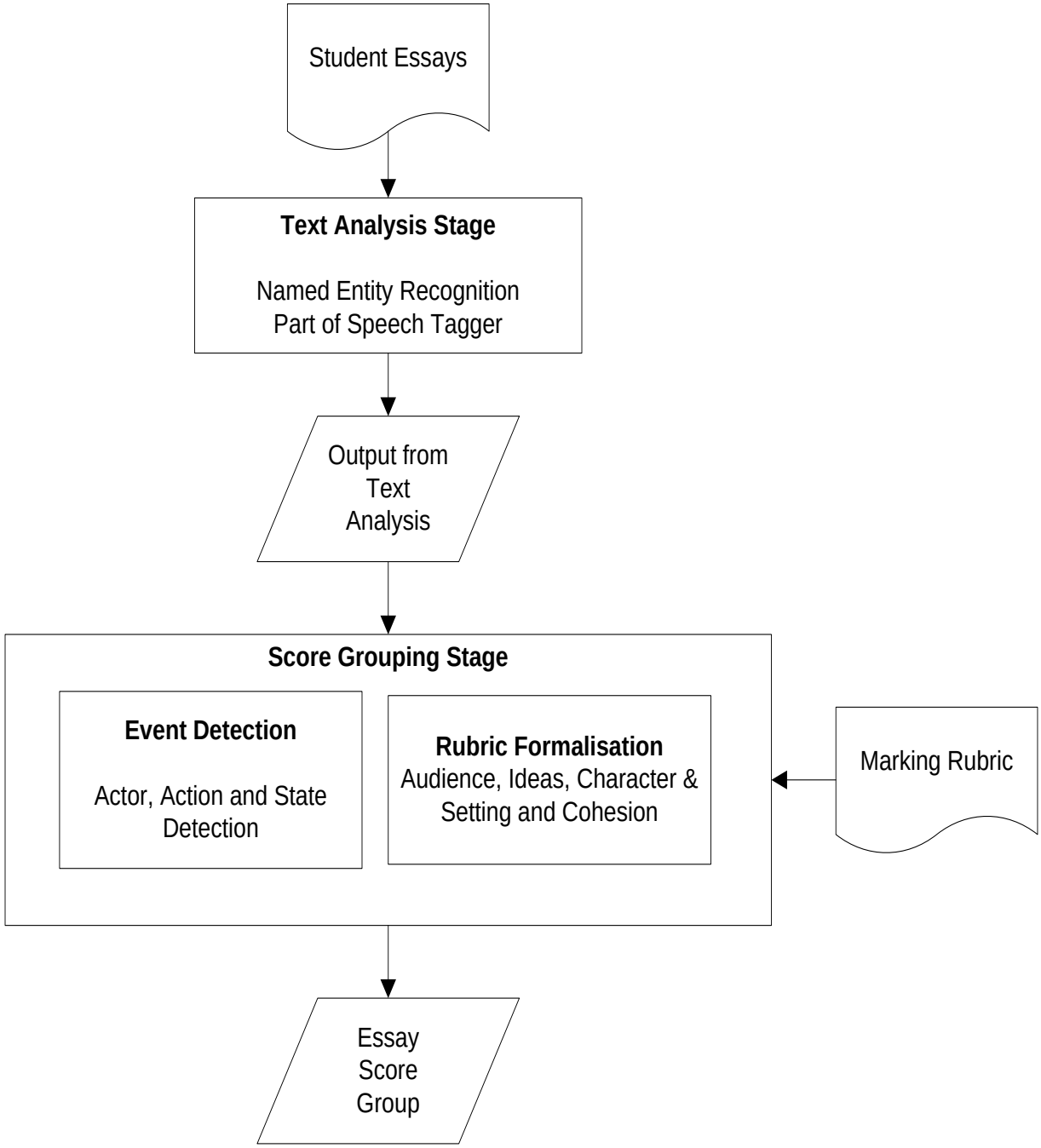

Figure 4.2: Overview of proposed solution

Since the use of Natural Language Processing tools plays a vital role in the development of the proposed solution, a detailed discussion of the two main tools used in this thesis are provided in the sections below.

4.3.2 Natural Language Processing tools

Of the many types of Human Language Technologies (HLT), Natural Language Processing (NLP) is probably the most complex. Unique from other forms of HLTs, such as Text Mining, Summarization or Generation, NLP is the practice of understanding the content of the text, rather than mainly focusing on extracting key pieces of information, like Text Mining.

Since some of the techniques that will be applied in this thesis will include certain tools used in NLP such as Part of Speech Tagging (POS) and Named Entity Recognition (NER), the next sections will briefly discuss these tools and their various applications in other works.

4.3.2.1 Part of Speech Taggers (POS)

The main aim of POS is to assist in recognizing patterns in natural language documents by automatically assigning tags (nouns, adjectives, verbs, etc) to words in the document's context, which facilitates more advanced analysis techniques in the text mining scope. Cutting et al. (1993) suggested a set of general requirements that POS applications should fulfil:

- The tagger should be able to distinguish between the actual body of the text and other isolated sentences such as the title, tables, references, etc. It should also be able to handle effectively words that are new to the tagger application.

- Any new data the tagger is exposed to should be 'learned' efficiently, requiring minimal computational time. The tagger should also be able to go through large corpuses of data efficiently.
- The tagger should have a minimal rate of error, in that it should correctly tag every word that it encounters.
- The tagger should be "tunable", in that it should be able to take in a human user's "insights" to avoid systematic errors.

Difficulties commonly faced in part of speech tagging are the lexical ambiguities that exist in most natural language documents. For instance, the words 'process' and 'programs' could be both tagged as either verbs or nouns, although this problem can be partially bypassed by analysing the context of the text itself. For example, in the abovementioned problem, if the word 'program' appears in a sentence as "...the program is part of a range of team building activities," it can only be a noun. Hence, the ambiguities of words are less of a problem when taken in the context of other words (Cutting et al. 1993).

One of the advantages of using POS is the ability to filter out non-significant words such as conjunctions and stop words, thus making the mining process more efficient.

Most part of speech tagging consists primarily of two steps (Daille 1994):

1. Term "candidates" are extracted based on the structure of the linguistic information, in other words, the context of the text. For example, candidates can

be selected based on morpho-syntactic patterns such as *noun-noun* (George Clooney) or *noun-preposition-noun* (Head of State).

2. Those candidates are then filtered according to one or more of a type of statistical relevance scoring scheme such as frequency of occurrence, similar information, log-like coefficients, etc.

One of the earlier POS applications was implemented by Brill (1992), in which a simple rule-based POS tagger was proposed. This was primarily designed to apply a predefined set of rules to the tagger which allows it to distinguish different parts of speech. Following the initial run on a particular training set, the tagger automatically recognizes its mistakes and attempts to correct them by implementing additional rules or 'patches' in the overall application logic. Though simple, the tagger performs on par with previous stochastic taggers and has a slightly reduced error rate with each new patch that is introduced into the tagger.

The main advantage of the rule-based tagger is that large stores of information are not required in order for the tagger to perform as well as other taggers that do, which also gives the rule-based tagger better portability from one corpus to another. Another advantage is that the tagger learns from its mistakes and automatically comes up with other 'rules' to rectify those mistakes.

Rajman and Besancon (1997) used four morpho-syntactic patterns (Noun-Noun, Noun of Noun, Adj-Noun and Adj-Verbal) to extract candidates, splitting the process up into the steps mentioned above to extract more complex compounds. For example, the

extraction of compounds such as “Oscar Winner George Clooney” first required the identification of two out of four predefined patterns (Adj-Verbal and Noun-Noun). The two patterns were then combined to form a unique compound which could be tagged as a Noun (‘Oscar Winner’ and ‘George Clooney’). Filtering was then done using a simple frequency based scheme.

4.3.2.2 Named Entity Recognition

Downey et al. (2006) define the process of Named Entity Recognition (NER) as the task of identifying and classifying names in textual documents. An alternative description is where NER is a sub-task of Information Extraction in which string elements are grouped into predefined categories such as persons, organisations or locations. In a more generalised explanation, Alfonseca and Manandhar (2002) state that NER involves the identification and classification of instances or objects of interest, which can fall under the above categories or “anything that is useful to solve a particular problem”.

In order to effectively and correctly extract information, Text Mining tools need to be able to distinguish which words or “linguistic constructions” represent “entities” (Witten 2003). Early NER tools used a set of rules that were input manually which, much like the problems faced in Brute Force type algorithms, require too much effort to correct and maintain. Modern methods of extracting entities are more inclined towards, though not limited to, the use of supervised methods in which an NER tool is first trained on a limited number of documents and the use of one of several machine learning techniques enables the tool to automatically decide which string elements constitute an entity.

4.3.2.2.1 Entity Types and Classifications

Entities are usually represented by more than one word but are seen as single vocabulary strings by NER tools (e.g. the name "Jane Smith" or the company "General Motors"). For example, consider the sentence, "Nokia was founded by Fredrik Idestam in Finland". Three named entities are present: 'Nokia' is an organization, 'Fredrik Idestam' is a person and 'Finland' is a location. The entities described above are those most commonly extracted by NER tools, generally termed "proper names" (Nadeau et al. 2006; Nadeau and Sekine 2007).

In the Message Understanding Conferences (MUC), the above named entities together with several others have been classified into three main expression types (Poibeau and Kossiem 2001):

- ENAMEX – Refers to proper names, e.g. persons, locations and organizations
- TIMEX – Refers to temporal expressions including dates and time
- NUMEX – Refers to numerical expressions such as money or percentages

Entity types are not limited to just the types described above. Fleischman and Hovy (2002) proposed a method in which the entity "person" could be further 'fine-grained' into eight subcategories which include "entertainer", "politician" and "businessperson". Previous works by Fleischman also split the entity "location" into several subcategories such as "City", "State", "Country", etc. (Fleischman 2001).

Other studies have proposed further breaking down entities into even more refined categories. Sekine et al. (2002) presented an extended hierarchy of named entities

which consisted of 150 categories. Simply speaking, the hierarchy is organized as a top-down tree structure where each category is broken down from a general class into more specific entity types. For example, an entity type 'Event' is further broken down into 'Games', 'Conference', 'Phenomena', 'War' and 'Natural Disasters'.

4.3.2.2.2 NER Learning Methods

Methods of recognizing named entities are known to fall under three general categories: supervised, semi-supervised and unsupervised learning, with supervised learning being the earliest and most widely-used method, and semi- and unsupervised learning being more recent developments. Nadeau and Sekine (2007) highlight this point, stating that in the 7th Message Understanding Conference (MUC), five out of the eight systems presented were based on supervised machine learning algorithms.

4.3.2.2.2.1 Supervised

As mentioned previously, supervised learning is the earliest (right after handcrafted rules) and preferred method of named entity recognition. The basic concept of supervised machine learning is to 'teach' the computer which instances represent entities by providing examples of positive and negative instances (Nadeau and Sekine 2007). In other words, the learning method is based on storing both right and wrong examples of named entities in a database to which the computer would refer when determining whether a set of words or strings represent a named entity.

A typical supervised machine learning method would first take in a large corpus containing a list of known entities and then attempt to identify other entities through a

set of rules, formulated by identifying distinguishing features of the known entities, and extracting string sets that share the same characteristics.

Examples of supervised learning methods include Decision Trees (Sekine 1998), Maximum Entropy (Borthwick 1998; Berger at al. 1996) and Hidden Markov Models (Bikel et al. 1999). Florian et al. (2003) proposed an NER method in which they combine four classifiers/algorithms namely: Robust Linear Classifier, Maximum Entropy, Transformation-based Learning and Hidden Markov Model. The resulting model was tested on the English and German languages and in the case of the English task, out-performed the best performing algorithms (Maximum Entropy and Robust Linear Classifier) by 17-21%. Performance on the German task yielded smaller improvement margins.

The main downside of this method is the large corpus of known entities which is required in order to allow the computer/machine to train itself, leading to problems due to the unavailability of such resources, or the high costs involved in acquiring them. As an alternative, semi- and unsupervised methods were developed to alleviate the burden of cost and also to reduce the amount of human intervention required.

4.3.2.2.2.2 Semi-Supervised

This method, as the name implies, sits in the middle between supervised and unsupervised learning wherein the main technique, known as “bootstrapping”, involves some level of supervision in which the algorithm is provided with a small sample of positive instances (Nadeau and Sekine 2007; Chapelle et al. 2006). In the context of

machine learning, bootstrapping is a method that progressively and iteratively improves performance by training and evaluating the recognition algorithm.

For example, an algorithm designed to recognize “Organizations” would first require the user to input a small number of examples. The algorithm would then search for sentences that contain those examples and try to recognise patterns within the context in which they occur. The algorithm would then attempt to identify other words or string sets that share the same contextual characteristics. This process is repeated iteratively on newly-found positive instances, allowing a large number of “Organization” type entities to be discovered. Riloff and Jones (1999) employ the above technique to automatically construct lexicon and extraction patterns, a process which they call “mutual bootstrapping”.

Collins and Singer (1999) proposed a named entity classification model that is based on the idea that an entity type can be easily distinguished by referring to both spelling and contextual rules; in other words, by looking at the spelling of the words and the context in which they appear. Supervision is reduced to a set of seven rules that the algorithm uses to extract candidate entities in a {spelling, context} format, which are then classified according to their context. Similar context characteristics in relation to the spelling of the word are then extracted to create a set of contextual rules, which in turn are used to identify other similar entities.

One of the main drawbacks associated with semi-supervised learning is that if the first examples or rules that are provided are incorrect or contain some form of ambiguity,

the subsequent learning of the algorithm would also be inaccurate or completely wrong. If this is the case, the algorithm would not yield any improved results and would in fact probably lower the performance by a large margin (Chapelle et al. 2006).

4.3.2.2.2.3 Unsupervised

Generally, unsupervised learning (ambiguity aside) involves the use of linguistic knowledge and lexical resources (e.g. WordNet) together with algorithms that deal with a large unannotated body of textual data. The more common approach to unsupervised learning in the field of NER is clustering, where entities are grouped together based on certain similarities such as the context in which they are used and where they appear. Unsupervised machine learning approaches in other fields include Quantile Estimation, Outlier Detection and Dimensionality Reduction (Nadeau and Sekine 2006; Chapelle et al. 2006).

An unsupervised method proposed by Nadeau et al. (2006) was able to indentify entities beyond the scope of general NER tools (e.g. the 3 classic entity types: Person, Location and Organization). The work borrows and expands concepts previously presented in Collins and Singer (1999) described in the above section and also from Etzioni et al. (2005) in which a large list of entities are generated for extraction. The first part of the system creates a large corpus of gazetteers of entities while the second part uses a set of heuristics founded upon previous works by Mikheev (1999), Petasis et al. (2001) and Palmer and Day (1997) to handle ambiguity between entities. The proposed NER tool was evaluated by comparing it with a baseline supervised method,

using the data used in the 7th MUC corpus. The results of the evaluation showed that the proposed method performed better than the baseline method although it fell short when compared to other, more in-depth systems.

4.3.2.2.3 Challenges with NER

4.3.2.2.3.1 Similarity between methods

There is usually some ambiguity between the classification of NER tools between the three types mentioned (Supervised, Semi-supervised and Unsupervised). Works such as those proposed by Collins and Singer (1999) present themselves as unsupervised stating that the tool requires little human intervention, such as a manually-annotated training set or hand crafted decision rules, although it could be argued that systems such as those cannot be truly considered unsupervised since the system still has a certain amount of reliance on human intervention, little as this may be.

In considering the distinction between the three types, the finest line is between semi-supervised and unsupervised methods. Nadeau et al. (2006) argue that in some works where systems appear to require no human labour, the generation of training sets is created just by “embedding clever rules and heuristics”. One salient fact is that systems described as unsupervised require considerably less supervision when compared to semi-supervised methods and also that the examples, when given, in unsupervised methods are usually unannotated or unlabelled.

4.3.2.2.3.2 Named-Entity Ambiguity

The main problems associated with NER tools centre around the ambiguity that occurs in most natural language documents, a problem that plagues most Text Mining tools. One simple reason this happens is because most lexical resources do not (either due to data storage constraints or pure impracticality) contain a complete dictionary of terms or definitions. This problem is even more prominent in recognizing named entities in the biomedical field since there exists an exponential volume of biological entities and each can be represented by more than one abbreviation or definition. Nadeau et al. (2006) listed three common ambiguity problems which are further explained below:

4.3.2.2.3.3 Entity-Noun Ambiguity

Ambiguity occurs when an entity and a noun share the same spelling but have different meanings; such pairs are called homographs. For example, the word "waters" can be either a surname or the plural of "water". One solution proposed by Mikheev (1999) uses a set of heuristics or rules that assumes that a word or phrase is a named entity where the initial letter is capitalized to be a Named-Entity unless:

a) The word or phrase sometimes appears without an initial capital letter

b) The word only appears at the start of a sentence or quotation

c) The word appears only in a sentence in which all words that have more than 3 letters start with a capitalized letter

4.3.2.2.3.4 Entity-Boundary Ambiguity

This is a common precision problem that occurs when a Named-Entity that is composed of more than one word is considered as two entities instead of one. For example, consider the following two phrases containing the names of organizations:

- "...organizations such as Apple and IBM..."
- "...such that Ernst and Young have more..."

In the first sentence, most NER tools are able to identify 'Apple' and 'IBM' as separate proper names. In the case of the second sentence, the NER tool might be unable to determine where the entity begins and ends, treating 'Ernst and Young' as two separate entities instead of one. Downey et al. (2005) refer to this as the "entity delimitation" problem.

Palmer and Day (1997) proposed the longest match strategy which Nadeau et al. (2006) used in a similar fashion. Their application of the solution involved a similar method in which all consecutive entities of the same type and entities with adjacent capitalized words are merged together, although entities of a different type were not merged since the resulting entity type would be lost or incorrect.

4.3.2.2.3.5 Entity-Entity Ambiguity

This problem occurs when the string that stands for a Named-Entity can belong to more than one type. An example given by Nadeau et al. (2006) is the string "France", which could either be the name of a person or the name of the country.

A solution that was proposed by Petasis et al. (2001), together with several others, was to take in the context of the string or word in question. For example, in the context Mr. France, the string "France" is in most cases a name instead of the country since it is preceded by the title "Mr." Other cues that can be used are those such as professional titles (e.g. Dr., Prof.) or organizational suffixes (e.g. Corp.).

4.3.2.2.3.6 Unseen Entity Class

Referred to as the "Entity Classes Problem" by Downey et al. (2006), the challenge for general NER tools when applied to Web applications is that the set of classes is not defined or known beforehand. Therefore, it is impractical to manually annotate each element of an entity class to provide a training set.

Downey et al. (2006) attempted to solve this problem by creating a training corpus where entities of any type are labelled as "entity", while negative examples of entities are labelled as such. This solution in itself proved to be slightly problematic since NER learning techniques are highly influenced by "orthographic and contextual features", both of which can vary widely across entity classes (Downey et al. 2006).

4.3.2.3 Summary of Part of Speech Taggers and Named Entity Recognizers

Although these tools are mainly applied to ad-hoc information retrieval tasks, their application to the field of essay grading is not uncommon. Most systems that utilise NLP processes incorporate these tools to perform more in-depth analysis of free text through their ability to break down the text into individual root forms. Even some

systems that employ statistical techniques such as word vectors need to identify parts of speech (nouns, verbs) before employing a particular algorithm.

Having provided a detailed discussion of Part of Speech taggers and NER tools, the next sections will continue to describe how those tools are incorporated into the proposed solution to provide the desired outputs for further analysis in the Group Scoring stage.

Firstly, the Text Analysis Stage where the raw text is pre-processed is described in detail. In order to perform this step, two Natural Language Processing tools are used, the Stanford Part of Speech tagger and Named Entity Recognizer.

Next, a description of the Score Grouping Stage, which is made up of two parts, namely the Event Detection Stage and Rubric Formalisation Stage is given; this is followed by a description of how Events are detected in the Event Detection Stage and finally, the Rubric Formalisation Stage is explained wherein the implicit details of the marking rubric are made explicit.

4.3.3 Text Analysis Stage for Essay Grading

This stage of the methodology deals with the processing of text into different outputs for further analysis. The first step in this process is to break down an essay into its individual sentences. The simplest way to do this is to refer to the punctuation within the essay. For example, full stops, question and exclamation marks are commonly used to denote the end of a sentence. This is done through a standard sentence splitter.

The next step involves parsing the text through the Stanford Named Entity Recognizer (NER), which provides an output of tags for every word in the sentence. The tags assigned to words are associated with proper name tags such as PERSON, LOCATION and TIME. Table 4.1 below provides some other examples of proper name tags.

Word	NER Tags
Susan	PERSON
England	LOCATION
12 Noon	TIME
Dollars	MONEY
SONY	ORGANIZATION
12	NUMBER

Table 4.1: Named Entity Recognition tags

The last step is to then tag each word with its relevant Part of Speech (POS) tag using the Stanford Part of Speech Tagger, the output of which would be used in conjunction with the NER tags to provide the output required to determine whether a sentence satisfies the criteria in order to be considered an Event. Table 4.2 below lists the most common POS tags and the words associated with them within a sentence.

Part of Speech Tags	Words in a Sentence
Determiner (DT)	***A*** normal day ***The*** empty house

Adjective (JJ)	The ***empty*** street A ***big*** dog
Noun (NN)	The empty ***house*** This sunny ***beach***
Personal Pronoun (PRP)	***He*** left home ***You*** left home
Possessive Pronoun (PRP$)	***His*** house was empty This house was ***mine***
Verb (VB)	He ***likes*** to eat He ***runs*** quickly
Adverb (RB)	He runs ***quickly*** He is **very** hungry
Coordinating Conjunction (CC)	Screaming ***and*** shouting ***But*** this time was different

Table 4.2: Common Part-of-Speech tags

It should be noted that some of these tags have certain variations. For example, the base POS tag for a verb is 'VB' but other instances of a verb can also be a verb in its past tense, 'VBD' or 'VBG' which is a present participle POS tag.

In some instances, it is of little importance which form of a word is tagged. For example, in looking for an Action, most times it is sufficient that the base tag be present; therefore, all variations of the POS tag for verb would be considered as its base tag. Other such assumptions, however, cannot be made for other POS tags such as

nouns, since in looking for an Actor, which is a character in the story, it is insufficient to assume that all nouns would refer to a character. This is further explained in the later sections.

Once the above steps have been performed under the Text Analysis stage, the Event Detection stage of the framework would then be performed using the output generated from the previous stage as its input.

4.3.4 Score Grouping Stage

This stage is primarily made up of two main parts:

- Event Detection and;
- Rubric Formalisation

The first is concerned with identifying those sentences in an essay which would qualify as an Event; the second determines which features of the essay relate to the descriptions given in the NAPLAN marking rubric.

The main goal of this stage is to combine the output of both parts in order to produce an accurate representation of how an essay would score in those respective criteria.

4.3.4.1 Event Detection

This part of the methodology involves using the output generated from the Text Analysis stage to determine whether or not specific sentences within an essay fulfil the Event criteria. For a sentence to be classified as such, there needs to be an:

- actor;

- action; and,
- state.

Therefore, the Event Detection stage needs to be able to identify, if present, each of these instances within each individual sentence. Figure 4.3 shows the overall process for Event Detection.

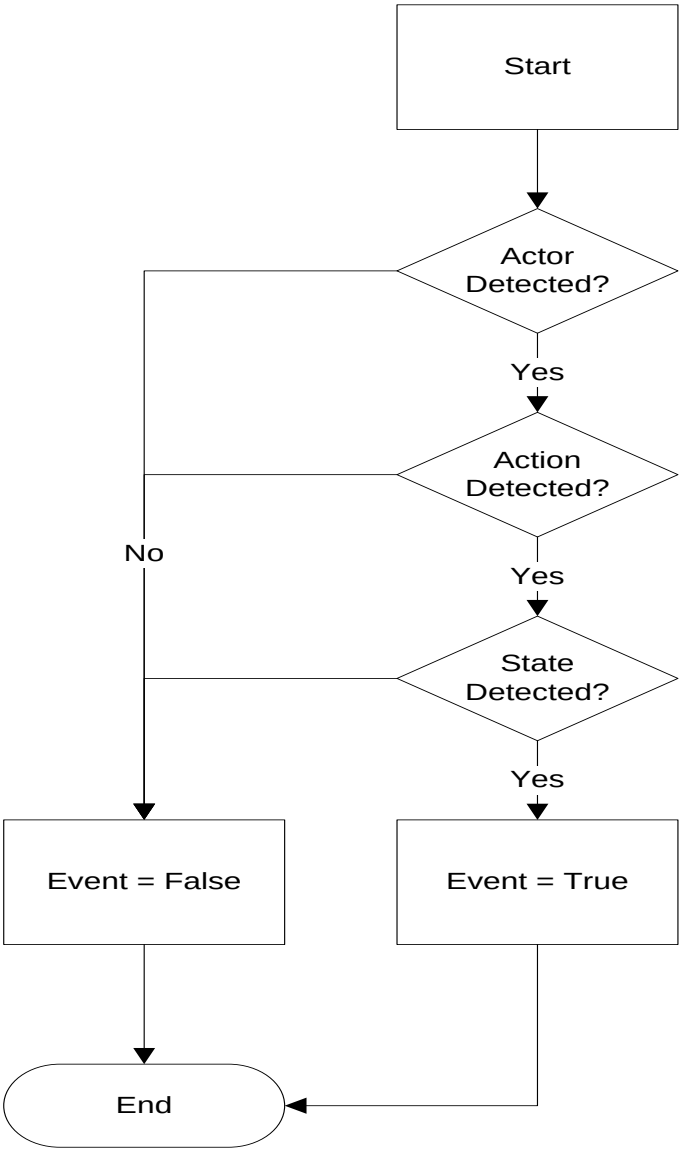

Figure 4.3: Overall process for Event Detection

4.3.4.1.1 Detecting Actors

As mentioned earlier, Actors within a story can be non-human characters. Therefore, the conventional search for proper nouns as denoted by NER tools, although accurate for the most part, might be insufficient. Relying solely on POS tags to pick out Actors is also unadvisable as not all inferences of a noun type POS tag necessarily refer to an Actor.

For example, a 'NN' tag might refer to the 'Police' or it might also refer to a character such as 'his father'. This issue can be addressed by adding other conditions and combining an NER tool with a POS tagger, making it possible to identify most Actors and at the same time disregarding whether or not they are human. Figure 4.4 below shows the process for the Actor detection method:

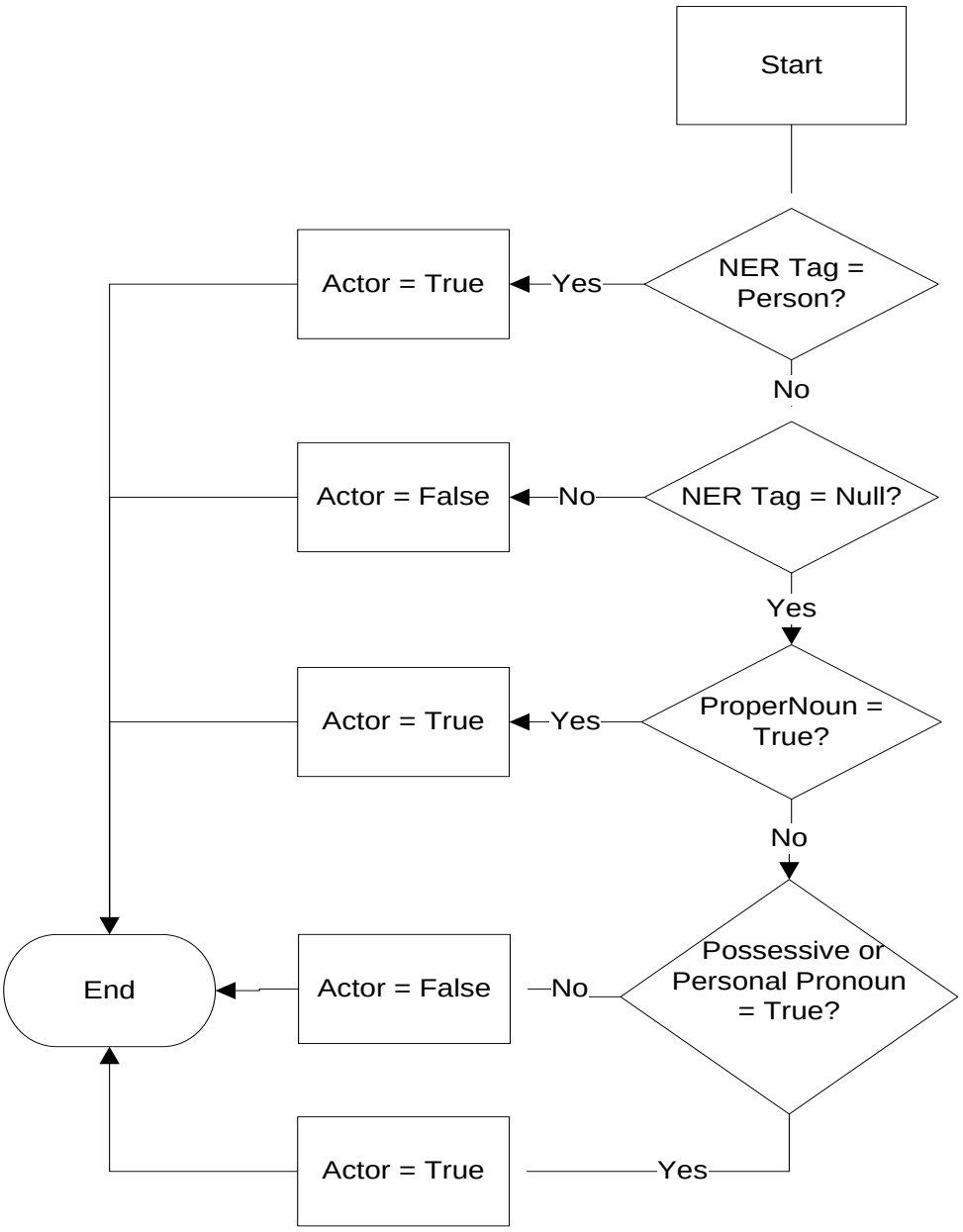

Figure 4.4: Process for detecting Actors

The first decision process does the initial check of the NER tags. If a word is tagged as PERSON, then there is no further need to perform any additional checking. Most PERSON tags are derived from proper name tags which relate to commonly-used names such as David or Susan. This serves as a sort of screening process to conserve where possible the amount of resources used in detecting Actors.

The second decision process checks if there is an NER tag attached to a word at all, since any other proper name NER tags would omit the possibility that the word can be considered an Actor.

The next decision process involves checking whether the POS tag of the word is a proper noun, tagged as 'NNP' by the POS tagger. Normally, if a word has the NER tag as 'PERSON' the POS tag would be 'NNP', but even without the NER tag, if this condition holds true, the word would still be assumed to be a character within the story and thus classified as an Actor.

The last decision checks whether the word is a possessive or personal pronoun. As mentioned earlier, the main objective is to find an Actor, which essentially is any character that is related to any Action or State that is detected within the sentence. Thus, if a word is a personal or possessive pronoun, it is assumed that it refers to a character within the story which indirectly means it can be classified as an Actor.

4.3.4.1.2 Detecting Actions

The process of detecting Actions for Event Detection purposes is, for the moment, a simplistic approach in that any word that is tagged as a verb (VB) is considered an Action. In this stage, only Part of Speech tagging is used.

However, to prevent Actions from interfering with pattern matching when checking for States, the following rule was applied:

"IF sentence length < 5 words, AND only one detected Action

THEN Action cannot be part of State pattern”

Figure 4.5 shows the process for detecting Actions:

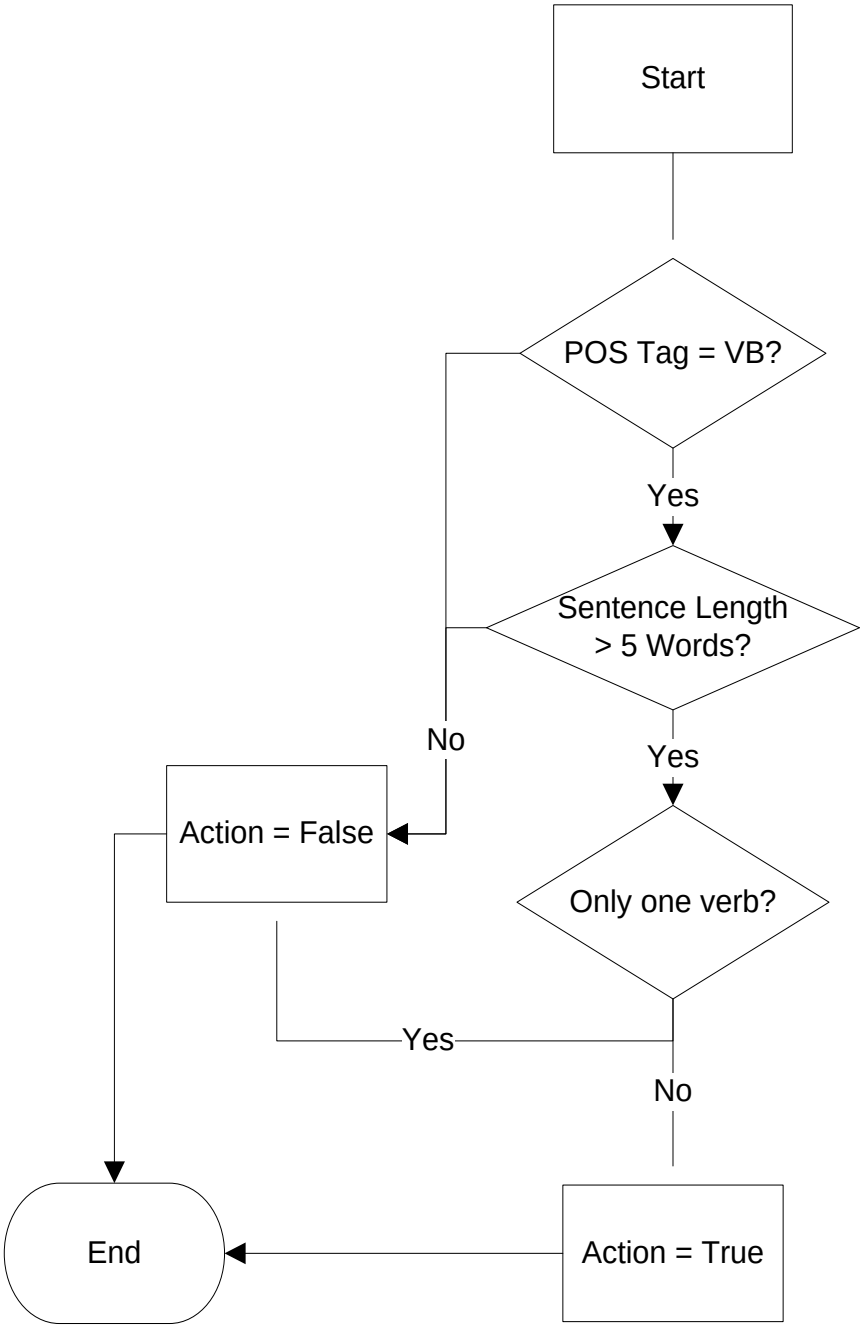

Figure 4.5: Process for detecting Actions

4.3.4.1.3 Detecting States

To reiterate, a State can be a time, location or condition. Furthermore, a condition could be either a mental or physical condition, such as being angry or tired. Picking out conditions could be done by brute force or having a predetermined list of common words together with their stemmed versions. For the purposes of this project, a lexicon of words is currently used to determine whether a mental or physical State exists within a sentence. The full list is shown in Appendix A.

Similarly, NER tools are relatively effective in determining a representation of time (12 noon, 3pm, 1 o'clock); therefore, there is little complication there. The challenge here is to detect a location.

Proper nouns such as London or Smallville that are recognised by the NER tool, are easily picked out, but narratives rarely, if at all, use such wording to describe locations. Oftentimes, locations would be "edge of the cliff" or "the rocky mountains". This is commonly referred to as the entity boundary problem or in some instances the unseen entity problem.

Therefore, pattern matching using the POS tagger is applied to pick out possible locations. Through human annotation of sample essays, locations were manually tagged in three-word patterns, then the most common patterns were identified which were then used to identify possible locations within the text, tagged as a Candidate Location (C. Location).

Consider the following phrases and their POS tags:

- In a cave - preposition (IN), determiner (DT), noun (NN)
- The jagged mountain – (DT), Adjective (JJ), (NN)

These are some of the more common patterns that could describe a possible Location. Table 4.3 below gives a full list of potential State (Location) patterns.

Patterns	Phrase
DT_JJ_NN	The sunny beach
IN_DT_NN	Into the room
IN_PRP$_NN	In his house
RB_IN_NN	Back from school
TO_DT_NN	To the house
VB_DT_NN	Entering the house
VB_PRP$_NN	Entered his house

Table 4.3: State (Location) Patterns

The patterns above are in fact Noun Phrases and can be found just by using a text parser to identify the groups of words which are Noun Phrases.

While this is certainly true, the difficulty in this is that while all Locations are nouns, not all nouns are Locations. Therefore, it is necessary for this step to be performed; using a pattern matching process, every such pattern is identified. Human intervention is again required to refine the accuracy of this technique.

If all three conditions are met, that is an Actor, Action and State are detected in a sentence, then it would be labelled as an Event.

Figure 4.6 below shows the basic process for determining if a State exists within a sentence.

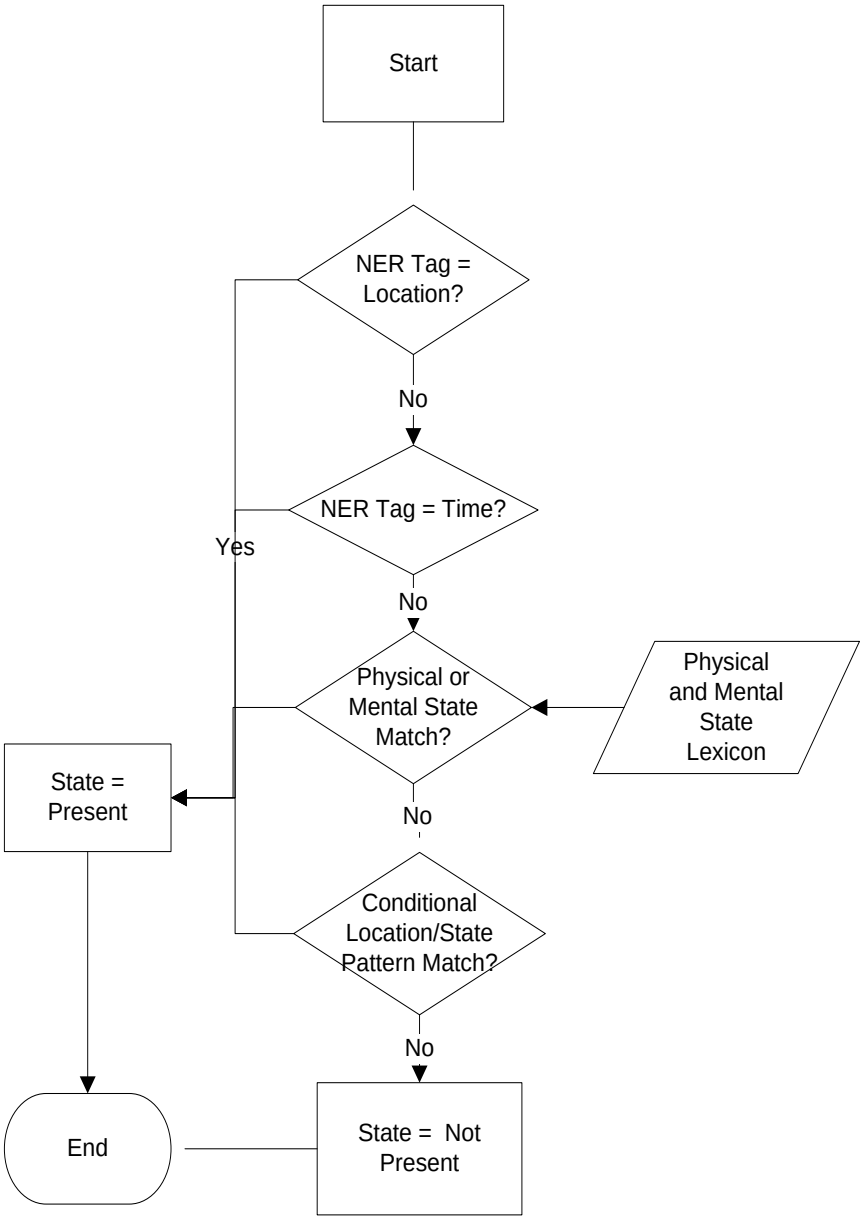

Figure 4.6: Process for detecting States

4.3.4.2 Rubric Formalisation

The NAPLAN marking rubric provides guidelines on how an essay should score with regards to ten categories with varying numbers of band scores. As mentioned earlier, an essay needs to fulfil certain conditions in order to receive the appropriate band score and some of these conditions are explicitly stated by the rubric. E.g. for the Paragraphing category, to receive a band score of 1, an essay must have at least two paragraphed sections of text.

For other categories, however, the rubric gives descriptions of how the text should be rather than providing explicit definitions. For example, for the Audience category, in order to receive a band score of 4 or more, the essay needs to "support[s] reader understanding and attempt[s] to engage reader". This is not surprising since the marking rubric was meant as a guide for human markers and not for machine translation.

Again, the four categories that will be focused on are:

- Audience –to orient, engage and affect the reader
- Ideas –creaton, selection and crafting of ideas
- Character & Setting –portrayal of character/development of a sense of place, time and atmostphere
- Cohesion –control of multiple threads and relationships

Also, as mentioned previously, each criterion in the NAPLAN rubric is further divided into a specific number of band scores which varies from one to the other. However, the

higher band scores of the criteria are often very hard to tell apart, often relying on the discretion and intuition of the human marker to determine, and even then there are disagreements.

Therefore, for the purposes of this thesis, we aim at broadening the Score Groups by splitting the band scores into three groups: Poor, Intermediate and Good. Table 4.4 describes the band scores of each criterion and the group under which they fall.

		Score Groups			
		0 Score	*Poor*	*Intermediate*	*Good*
Marking Criteria	*Audience*	0	1-3	4	5-6
	Ideas	0	1-2	3	4-5
	Character & Setting	0	1-2	3	4
	Cohesion	0	1-2	3	4

Table 4.4: Score Groups according to Band Scores

Apart from these three main groups, an essay can fall under the 0 Score group. Simply put, none of the content of such an essay meets any of the criteria.

The reason for this 0 Score grouping is to filter out all those essays which do not need further analysis, thus reducing the amount of resources required and shortening the processing time. This filtering process is not discussed here in detail but is instead covered in another complementary work being carried out in conjunction with this thesis, which deals with the more technical aspects of a narrative essay. However, to give a better understanding of how each scoring logic works, the 0 score grouping is mentioned as a step in each of the scoring logics.

A two-step process is used to determine the score group of an essay. First, we need to define explicitly which features correspond to what is perceived to be adequate in order to achieve a particular band score. These features would be weighed according to their significance in relation to that particular criterion. For example, in the case of Ideas, should the main condition for a good score be that the essay has a good ratio of Events to non-Events, an essay that has a ratio of between 35% and 85% would therefore receive more points than one that falls outside of this ratio.

The second step takes into account the certain specified conditions that place an essay in one of the three categories namely A, B or C, with C being the poorest. For example, under the Ideas criterion, if an essay has a ratio between 35% and 85% and is longer than 30 sentences with a high number of descriptive words, it would be placed in category (CAT) A. Once these steps have been performed, the output from both these processes would be combined and used to determine the score group to which an essay belongs.

The next sections describe in detail the score grouping processes for each criterion, first explaining how features (Event ratio, essay length, etc.) would contribute to its weighted score; secondly, how the presence or absence of certain conditions would place an essay into one of the three aforementioned categories is described; and finally, how these processes are combined to place an essay in its relative score group.

4.3.4.3 Audience

This criterion is made up of seven band scores, numbering 0 to 6. The general description relates to the author's attempt to involve the reader, in this case with regards to containing sufficient information to allow the reader to properly follow the story. First off, if an essay has no sentences or contains only some symbols or drawings, it would fall within the 0 Score group.

4.3.4.3.1 Scoring

In determining the score an essay receives under this criterion, the features that are taken into consideration are:

- Essay length
- Number of Events
- Event Ratio

Each essay is assigned a base score of 5, with the values of the above features adding to or subtracting from that score. The process for carrying this out is displayed in JAVA code in Table 4.5 below. For the full source code, refer to Appendix B.

```
//check number of Events and Ratio
if (noOfEvents>1)
                  {if (ratio >= 0.35 && ratio <= 0.85)
                              {if (noOfEvents> 15)
                                          score = score + 0;
                                    elseif (noOfEvents>13 )
                                          score = score - 1.5;

                                    elseif (noOfEvents>= 10)

                                          score = score - 2.5;

                                    elseif (noOfEvents>= 8)
                                          score = score - 3;
```

```
                                    elseif (noOfEvents>= 5)
                                          score = score - 3.5;
                                    else
                                          score = score - 5;}
                              else {//double check Events
                                    if (noOfEvents>= 15)

                                          score = score - 1

                                    elseif (noOfEvents>=12 )
                                          score = score - 2;

                                    elseif (noOfEvents>= 10)

                                          score = score - 2.5;

                                    elseif (noOfEvents>= 8)
                                          score = score - 3.5;
                                    elseif (noOfEvents>= 5)
                                          score = score - 5.5;
                                    else
                                          score = score - 6;}

                  }else score = score - 10;
//check Essay Lengthif(essayLength> 1)
{if (essayLength> 30)
                              score = score + 0;

                        elseif (essayLength>= 25)
                              score = score - 2;
                        elseif (essayLength>15)
                              score = score - 2.5;
                        elseif (essayLength> 9)
                              score = score -3.5;
                        else
                              score = score - 5;}
                  else
                        score = score - 8;
```

Table 4.5: Audience Score grouping source code

4.3.4.3.2 Grouping

In formalising this criterion, the essay length (number of sentences) and the presence of Events would contribute to an essay's score, since a relatively short essay with fewer than twenty-five sentences would not have sufficient information or elaboration to allow for a higher band score. Therefore, if an essay has fewer than twenty-five

sentences, it would be placed no higher than CAT B. Should an essay contain a very short script, which is less than eight sentences, then it can only be placed in CAT C.

Since an Event can be seen as a story device, it can be assumed that without any Events, there is little attempt to engage or involve the reader. Therefore, if none are detected, the essay would also be placed in CAT C.

To be placed in CAT A, the essay should include descriptions of the emotional or physical conditions of the Actors. The emotional conditions detected for an Actor would thus add to an essay's score under this criterion. Therefore, in addition to the essay features used in the scoring phase, the presence of a physical or mental state will also be considered.

Figure 4.7 depicts the grouping logic for the Audience criterion.

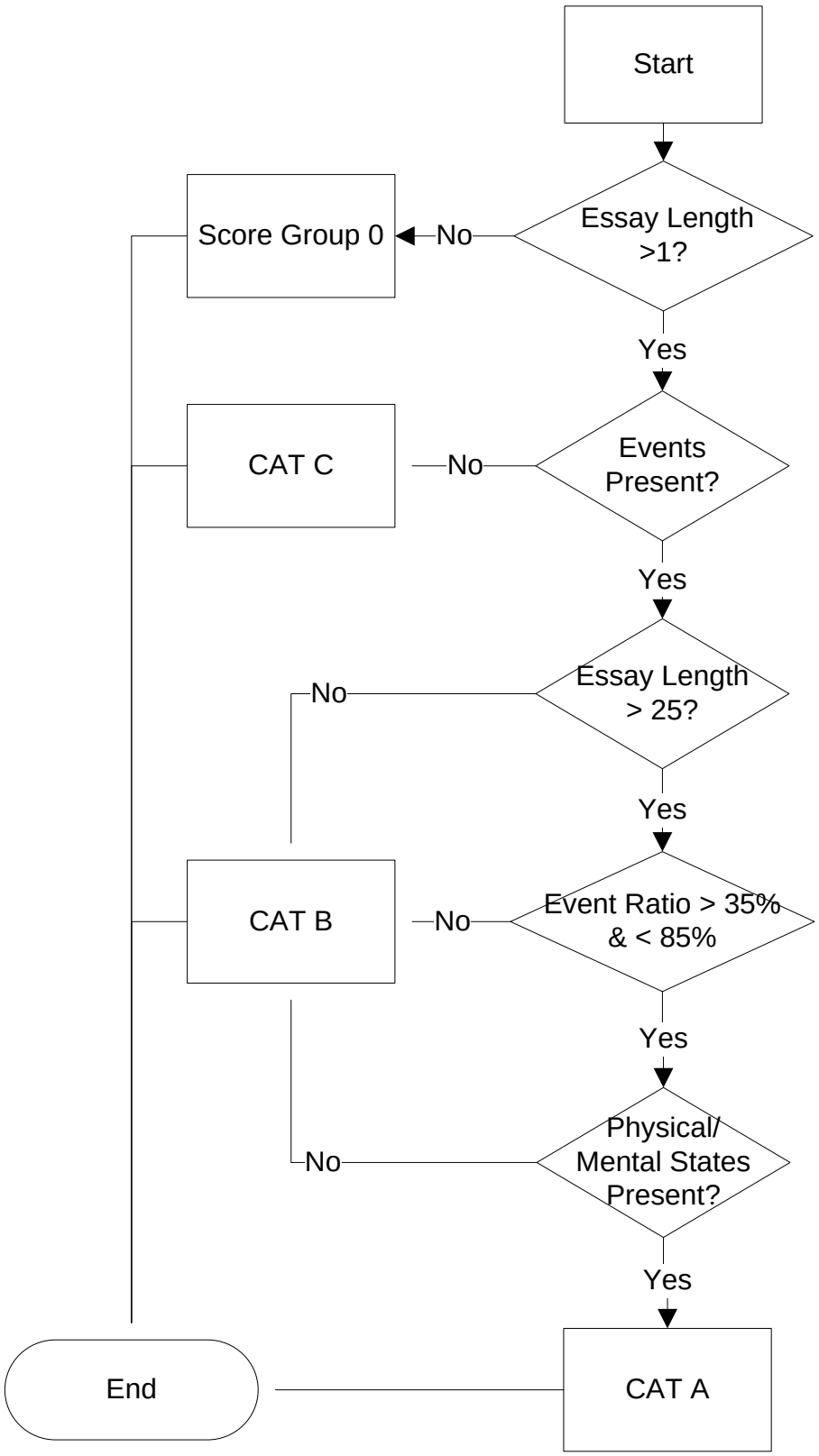

Figure 4.7: Grouping logic for Audience

4.3.4.3.4 Score Grouping

Once the two steps above have been performed, the resulting output is combined and used to determine the Score Group to which an essay should belong. Figure 4.8 describes the process:

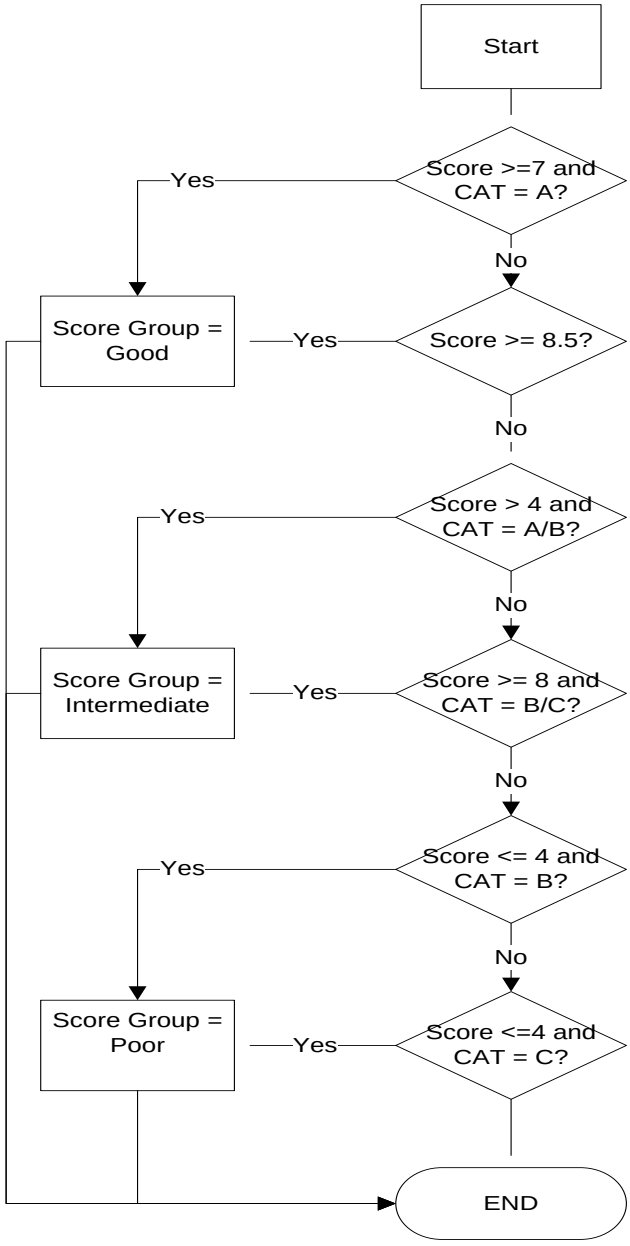

Figure 4.8: Score Grouping logic for Audience

4.3.4.4 Ideas

According to the NAPLAN marking rubric, the Ideas criterion is split into six band scores, numbering 0 to 5. The formalisation for the Ideas criterion is relatively straightforward in terms of defining what to look for. In the proposed solution, the identification of an Event is taken to mean that the author has managed to convey an 'idea'.

It is therefore theorised that in identifying all the Events in the essay, together with calculating the ratio of Events with regards to the total number of sentences, this would enable an appropriate band score to be assigned.

The detection of an Event within the text signifies the presence of a character (Actor), happenings (Actions) observed or performed by the character and the situation (State) of the character. Hence, the presence of an Event signifies that an 'idea' has been created and expressed by the author.

Determining if an essay receives a score of 0 is relatively easy since such an essay would theoretically contain no Events at all. Alternatively, it could also mean that the essay is in fact made up of illustrations or figures that attempt to convey meaning, in which case the essay would still receive a score of 0.

4.3.4.4.1 Scoring

The features of an essay considered for this criterion are as follows:

- Essay length
- Number of Events
- Event Ratio

- Number of unique adjectves
- Number of unique adverbs

The presence of a high number of adjectives and adverbs, while not directly related to how well an essay is written, does however indicate a more descriptive story and therefore more descriptive Events. Hence, the presence of a high number of unique adverbs and adjectives would add to the score the essay receives in this respect. The logic for scoring under this criterion is shown in Table 4.6:

```
//check number of Events and Ratio
                  if (noOfEvents>1)
                  {
                              if (ratio >= 0.35 && ratio <= 0.85)
                              {
                                    if (noOfEvents> 15)
                                          score = score + 0;
                                    elseif (noOfEvents>13 )
                                          score = score - 1.5;

                                    elseif (noOfEvents>= 10)

                                          score = score - 2.5;

                                    elseif (noOfEvents>= 8)
                                          score = score - 3;

                                    elseif (noOfEvents>= 5)
                                          score = score - 3.5;
                                    else
                                          score = score - 5;
                              }
                              else
                              {
                                    //double check Events
                                    if (noOfEvents>= 15)

                                          score = score - 1;
                                    elseif (noOfEvents>=12 )

                                          score = score - 2;

                                    elseif (noOfEvents>= 10)
                                          score = score - 2.5;

                                    elseif (noOfEvents>= 8)
                                          score = score - 3.5;
```

```
                        elseif (noOfEvents>= 5)
                              score = score - 5.5;
                        else
                              score = score - 6;

                  }


            }

            else score = score - 10;

            //check Essay Length
            if(essayLength> 1){
                  if (essayLength> 30)
                        score = score + 0;

                  elseif (essayLength>= 25)
                        score = score - 2;
                  elseif (essayLength>15)
                        score = score - 2.5;
                  elseif (essayLength> 9)
                        score = score -3.5;

                  else

                        score = score - 5;

            }
            else
                  score = score - 8;

            //check number of adjectives
            if(adj> 1){
                  if(adj>=20)
                        score =score + 0;
                  elseif (adj> 15)
                        score = score - 1;

                  elseif (adj> 10)
                        score = score - 1.5;


            }

            else

                  score = score - 2;



            //check number of adverbs
            if(adv > 1){
                  if(adv >=20)
                        score =score + 0;
                  elseif (adv> 15)
                        score = score - 1;
```

```
                elseif (adv> 10)
                        score = score - 1.5;

            }
            else
                        score = score - 2;
```

Table 4.6: Ideas Score grouping source code

4.3.4.4.2 Grouping

For an essay to be within CAT C it would need to contain very few Events and the ratio of Events to non-Events would be rather extreme (for example 0 or 100 percent). A short script would also be classified as belonging to CAT C. Hence, if an essay's length is less than eight sentences, it would not qualify for a better category.

However, essays within CAT B would contain more than eight sentences, with a good mix of Events and non-Events. This mix is determined by the ratio of Events and non-Events over the total number of sentences within the essay.

Similarly, for an essay to be within CAT A, with regards to the Ideas criterion, it would need to have the same characteristics of a CAT B essay but have a longer script including a more ideal ratio of Events and non-Events. The logic is shown in Figure 4.9 followed by the Score grouping logic in Figure 4.10.

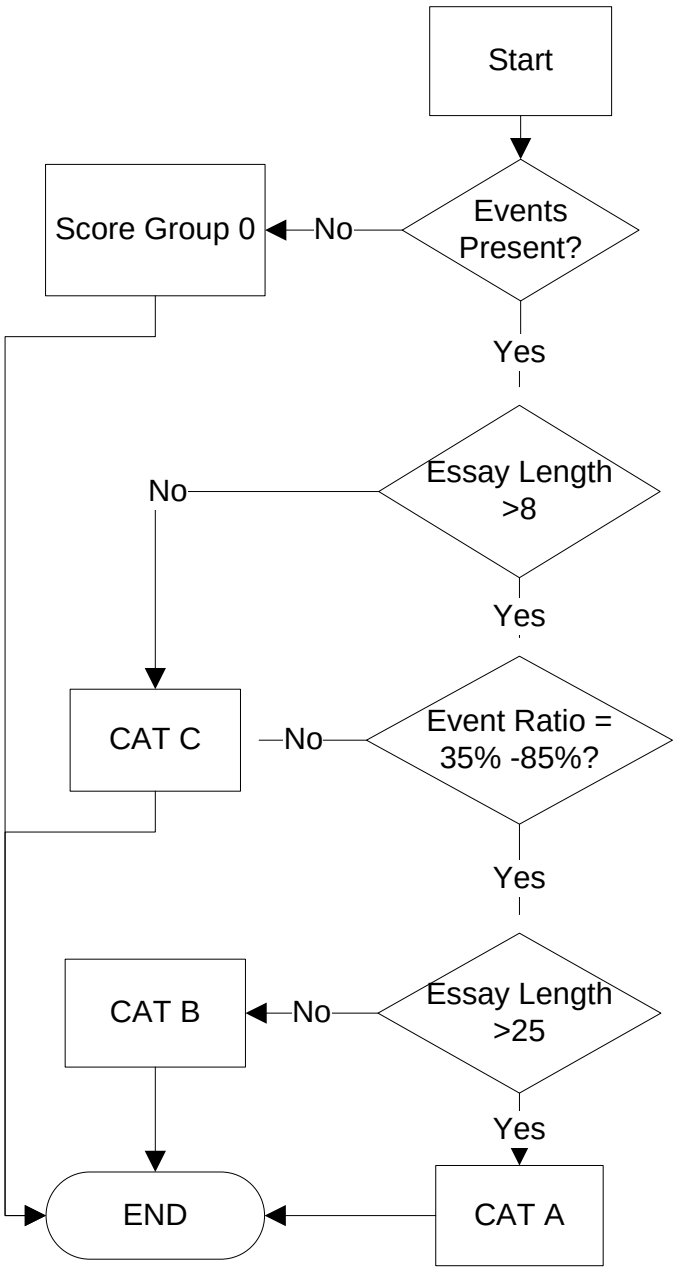

Figure 4.9: Ideas grouping logic

4.3.4.4.3 Score Grouping

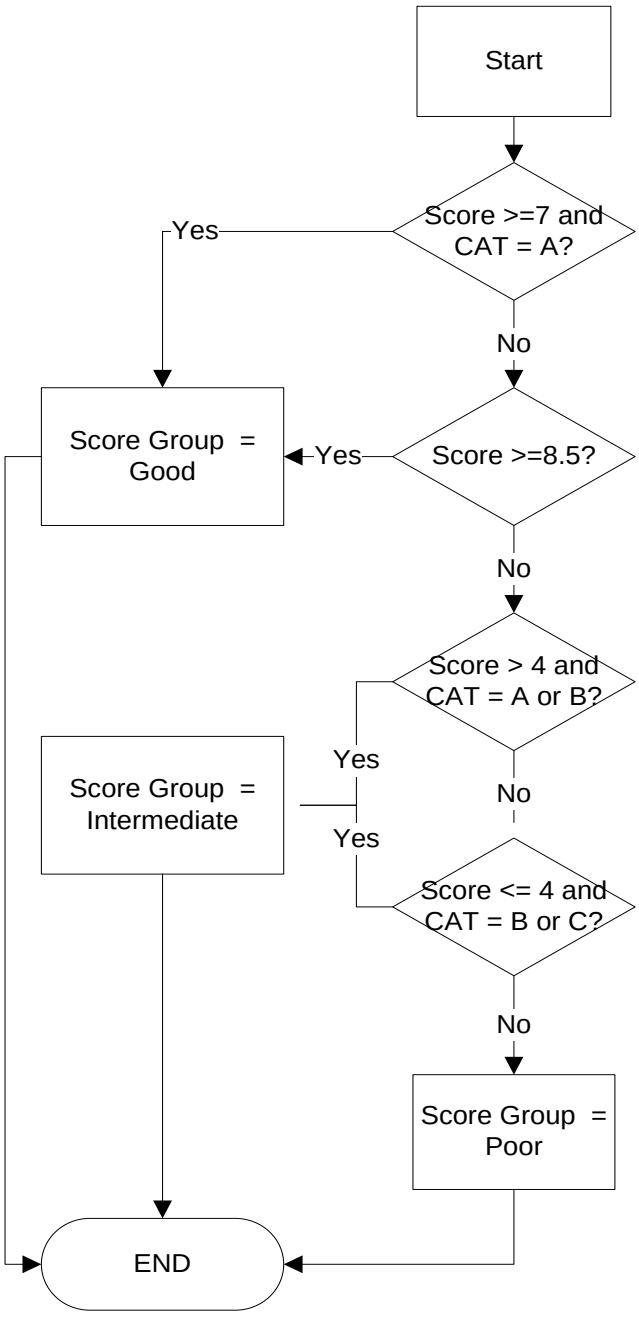

Figure 4.10: Ideas Score grouping logic

4.3.4.5 Character and Setting

This criterion is assumed to be linked somewhat to the Ideas criterion since its score also relies on the presence of Events within the essay. The difference here is that the focus is more on the development of the characters within and/or the setting (State) depicted.

It should be noted that according to the marking rubric, Character and Setting is listed as an 'either or' condition, meaning that an elaboration of either one (Actors or State) is sufficient.

This criterion should be rather simple since we can take the conditions for scoring Ideas and apply them to this criterion albeit with some modifications regarding the focus. For example, assigning a 0 score would need some adjustments since, even if no sentences in the essay are determined to be Events, an essay would not receive a 0 score as long as an Actor or State is detected.

4.3.4.5.1 Scoring

Similar to the Ideas criterion, the features that are used in this scoring process are the:

- Essay length
- Number of Events
- Event Ratio
- Number of unique adjectives
- Number of unique adverbs

Table 4.7 describes the logic, shown in JAVA code:

```
//Check Ratio and Events
                  if (noOfEvents>1)
                  {
                              if (ratio >= 0.35 && ratio <= 0.85)
                              {
                                    if (noOfEvents> 15)
                                          score = score + 0;
                                    elseif (noOfEvents>13 )
                                          score = score - 0.5;

                                    elseif (noOfEvents>= 10)

                                          score = score - 1;

                                    elseif (noOfEvents>= 8)
                                          score = score - 1.5;

                                    elseif (noOfEvents>= 5)
                                          score = score - 3.5;

                                    else

                                          score = score - 5;

                              }
                              else
                              {
                                    //double check Events
                                    if (noOfEvents>= 1)
                                          {     if (noOfEvents> 15)
                                                      score = score +
1;
                                                elseif (noOfEvents> 10)
                                                      score = score +
0.5;
                                                elseif (noOfEvents> 5)
                                                      score = score -
4;
                                                else
                                                      score = score -5;

                                          }

                                          else

                                                score = score - 8;

                              }
                  }
                  else score = score - 10;
                  //check Essay Length
                  if (essayLength>= 30)
                        score = score + 1;
                  elseif (essayLength> 24)
```

```
            score = score + 0.5;
      elseif (essayLength> 14)
            score = score - 0.5;
      elseif (essayLength> 9)
            score = score - 1;
      else
            score = 0;

      //check number of adjectives
      if (adj>=1)
            {if (adj> 20)
                  score = score + 1;
            elseif (adj> 15)
                  score = score + 0.5;
            elseif (adj>= 10)
                  score = score - 1;

            elseif (adj > 5)

                  score = score - 4;

            }
      else
            score = score - 8;
      //check number of adverbs
      if(adv > 1){
            if(adv >=20)
                  score =score + 1;
            elseif (adv>= 15)
                  score = score + 0.5;

            elseif (adv>= 10)
                  score = score + 0;
            elseif (adv>= 6)
                  score = score - 2;
            else
                  score = score - 6;
      }
      else
            score = score - 8;
```

Table 4.7: Character and Setting Score grouping source code

4.3.4.5.2 Grouping

If, within the essay there is mention of Actors only as names (Daniel, Susan, etc.) or roles (father, mother, etc.) or Settings (simple locations such as 'the beach' or 'school') then it would be classified as CAT C.

In order to achieve a higher grouping, these aspects of the essay require greater description or elaboration. This might translate into descriptions of Actors which are implied by the use of adjectives. Furthermore, descriptions of Locations which might also be represented by the use of adjectives such as ‘sunny’ or ‘windy’ would add to the elaboration of the setting and hence place an essay in CAT B.

An essay in CAT A should be highly detailed with characteristics being given to the Actors, or the current State being described well. This would theoretically mean that the essay should have a higher number of adjectives describing the State and Actors and also a high number of adverbs in relation to the Actions performed. Furthermore, an attempt to depict the situation or condition of a character should be detected; hence, in addition to the above requirements, a physical or mental State should be present.

Figure 4.11 and 4.12 illustrates the scoring logic and group scoring respectively:

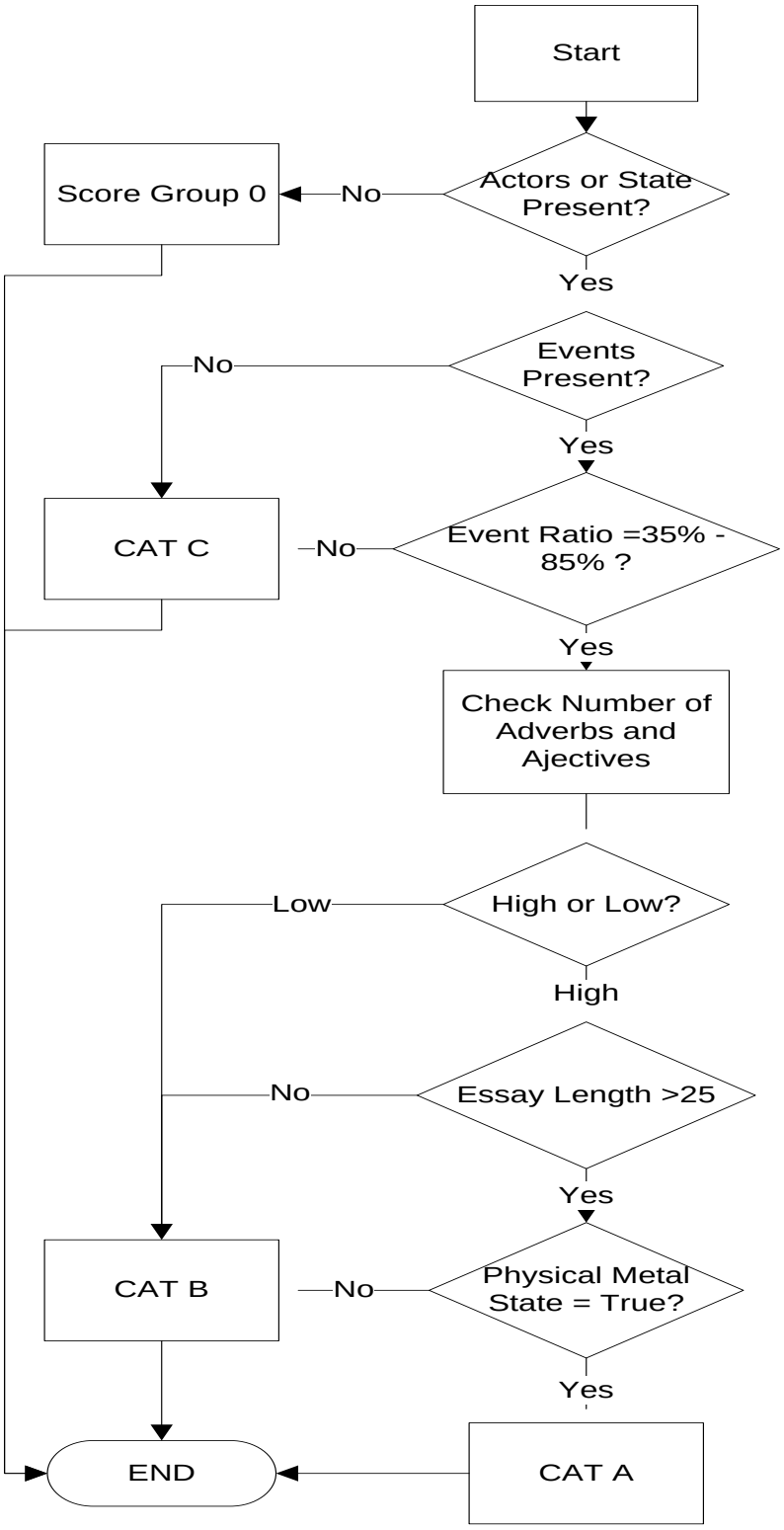

Figure 4.11: Character and Setting grouping logic

4.3.4.5.3 Score Grouping

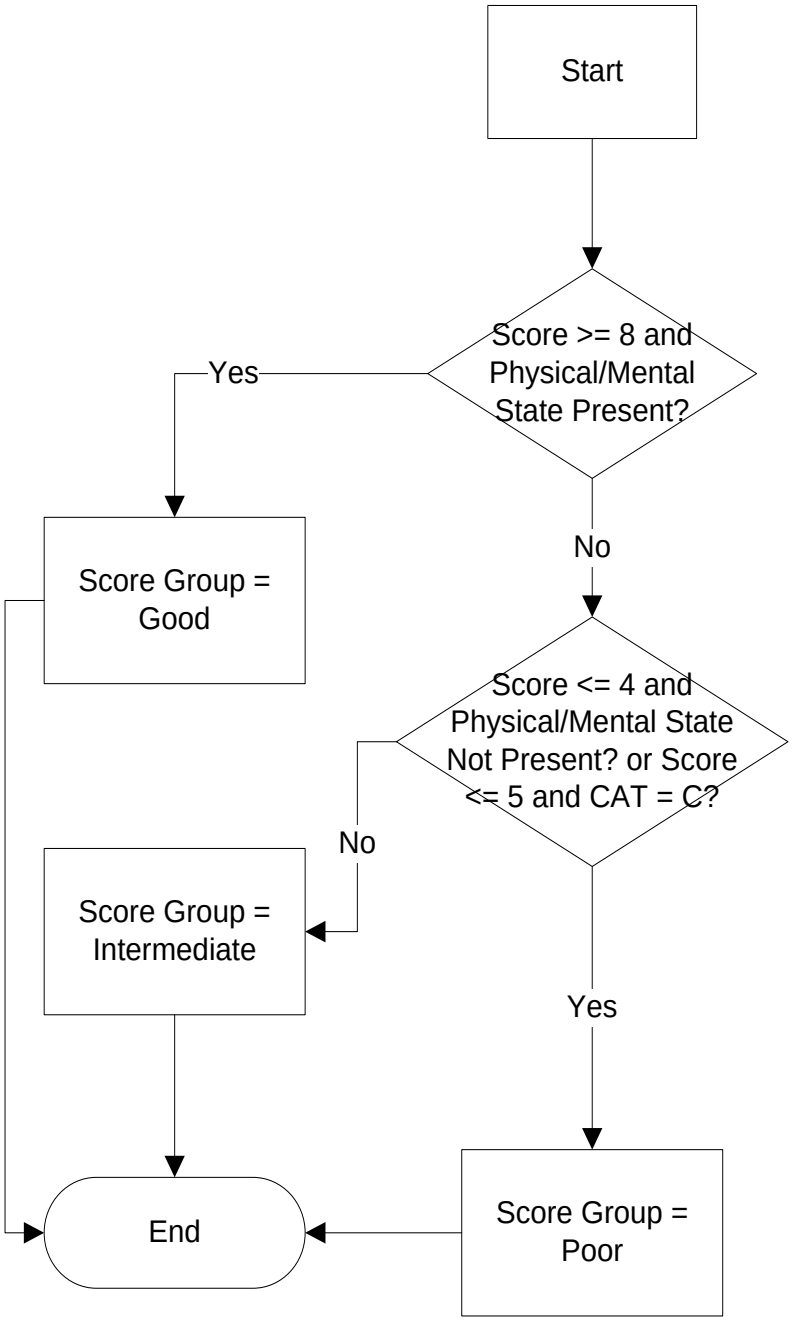

Figure 4.12" Character and Setting Score grouping logic

4.3.4.6 Cohesion

The goal of this criterion is to determine whether the essay flows in an appropriate manner. This can be done by looking at the author's use of referring words, substitutions, word associations and connectives. For example, some simple connectives may be 'then', 'soon', 'and' etc. Also, there should be a variety of these connectives and not just ones that are used repeatedly. For the full list of connectives that are checked for, refer to Appendix C.

4.3.4.6.1 Scoring

The features used in determining the score an essay receives for Cohesion are as follows:

- Essay length
- Number of Events
- Event Ratio
- Number of simple connectives
- Number of advanced connectives

Table 4.8 shows the scoring process:

```
//Check number of simple connectives (4)
                  if (simpleCon != 0)
                  {
                        if (simpleCon>=10)
                              score = score + 3;
                        elseif (simpleCon>= 5)
                              score = score +1.5;

                        elseif (simpleCon>= 1)

                              score = score + 0.5;
```

```
                    }
                    else
                          score = score + 0;

                    //Check Number of Advanced Connectives (2)
                    if (advCon>=1)
                          score = score + 2;

                    //Check Ratio and Events
                    if (noOfEvents>1)
                    {
                                if (ratio >= 0.30 && ratio <= 0.39 ||
ratio >= 0.60 && ratio <= 0.85 || ratio > 0.50 && ratio < 0.59)
                                {
                                      if (noOfEvents> 15)
                                            score = score + 0;
                                      elseif (noOfEvents>13 )

                                            score = score - 0.5;

                                      elseif (noOfEvents>= 10)

                                            score = score - 1;

                                      elseif (noOfEvents>= 8)

                                            score = score - 1.5;

                                      elseif (noOfEvents>= 5)

                                            score = score - 3.5;
                                      else
                                            score = score - 5;
                                }
                                else
                                {
                                      //double check Events
                                      if (noOfEvents>= 1)
                                            {     if (noOfEvents> 15)
                                                        score = score +
1;

                                                  elseif (noOfEvents> 10)
                                                        score = score +
0.5;

                                                  elseif (noOfEvents> 5)
                                                        score = score -
4;


                                                  else

                                                        score = score -5;

                                            }
                                            else
                                                  score = score - 8;
                                }
                    }
```

```
else score = score - 10;

//check Essay Length
if (essayLength>= 30)
      score = score + 1;
elseif (essayLength> 24)
      score = score + 0.5;
elseif (essayLength> 14)
      score = score - 0.5;
elseif (essayLength> 9)
      score = score - 1;
else
      score = 0;
```

Table 4.8: Cohesion Score Grouping source code

4.3.4.6.2 Grouping

Similar to other criteria, if an essay contains only symbols or drawings that attempt to convey meaning but in fact has no legible sentences or words, it would be placed in the 0 Score group.

Therefore, as long as there is some content, if even one of these connectives or referring words is found within a sentence of the essay, the essay would be placed in CAT C but no higher without other requirements being fulfilled.

Alternatively, an essay that uses a high number of connectives such as ‘meanwhile’ or ‘concurrently’ together with a low repetition rate would be placed in a higher category. The difference that separates an essay in CAT A from one in B is the type and number of connectives that it contains.

An essay that uses advanced connectives together with simple ones is classified as having a high level of Cohesion and is placed in CAT A. Figure 4.13 illustrates this:

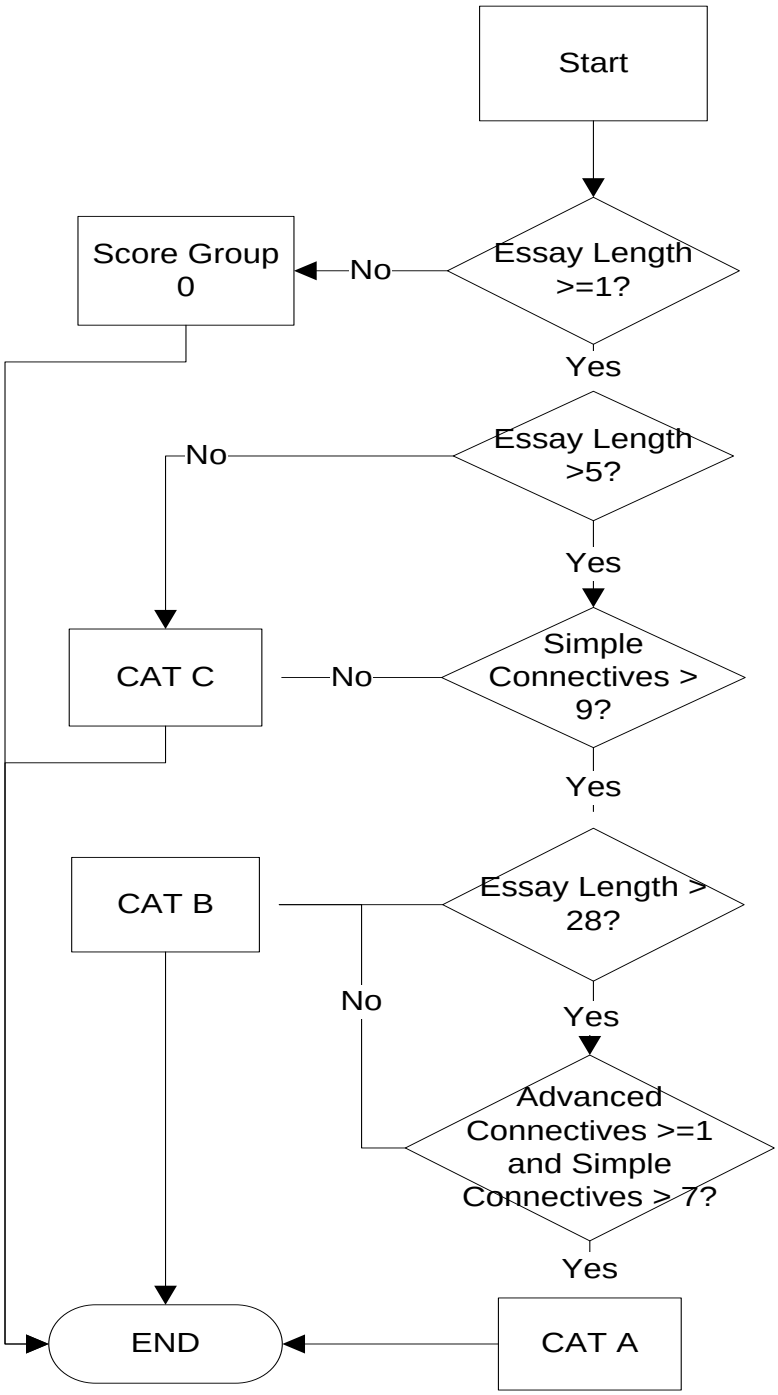

Figure 4.13: Cohesion grouping logic

4.3.4.6.3 Score Grouping

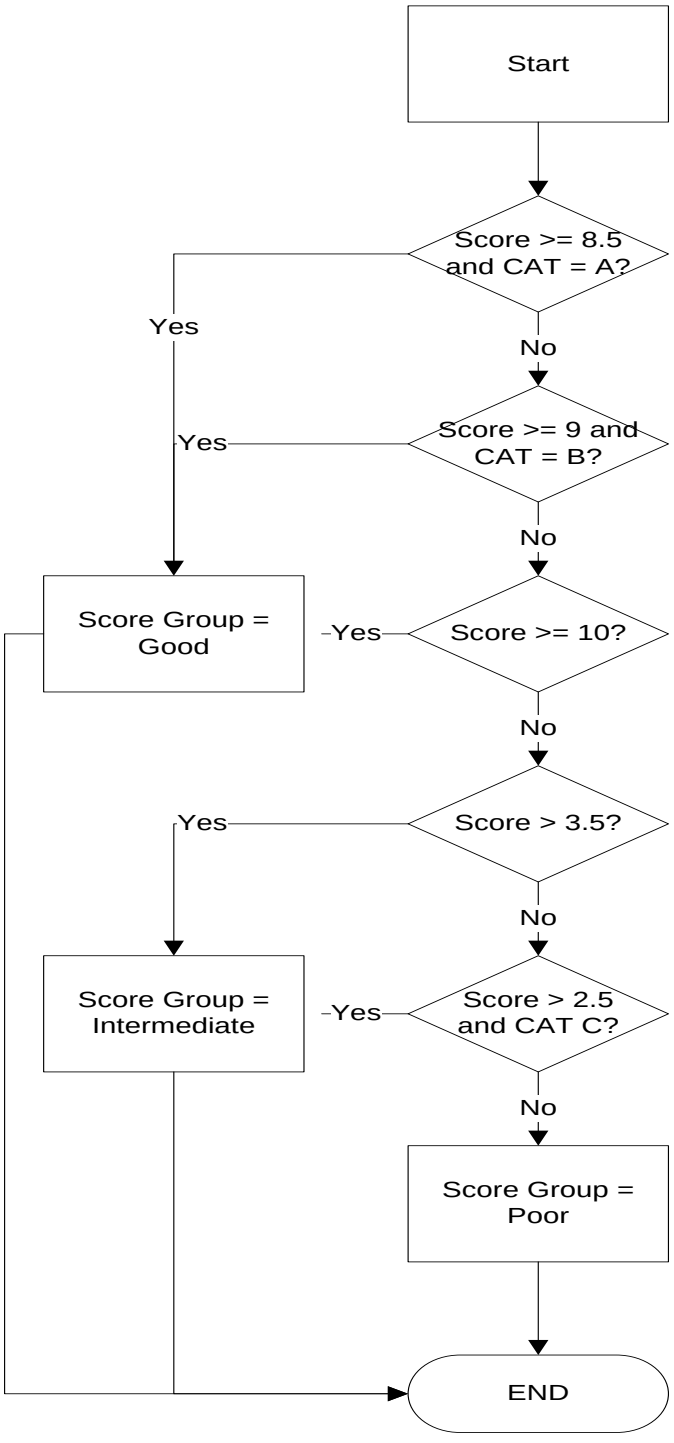

Figure 4.14: Cohesion Score grouping logic

4.4 Conclusion

In this chapter, the conceptual framework was presented together with an explanation of each individual stage and its various sub-stages. Events are detected through a series of existence checks performed on the output generated by the Text Analysis Stage. In turn, the output from the Event Detection step is combined with the scoring logic based on the formalisation of the NAPLAN marking rubric.

While many methods and frameworks have been developed for the grading of essays, the way in which the proposed solution differs from the rest is that it does not rely heavily on a knowledge base. This means that the system requires no further training once it has been initially performed and it can be applied to any writing genre.

While it may appear that it performs only what can be seen as an analysis of the text, the output that is generated by those analysis stages is sufficient for the aims of the framework. The Event Detection framework is thus able to provide the necessary output to group an essay according to how it performs under the aforementioned criteria without having to undertake resource-hungry processes.

The next chapter describes in detail the Event Detection process, performed on a test bed of sample student essays.

Chapter 5 - Detailing the Event Detection Process

5.1 Introduction

With the conceptual framework of the proposed solution explained in Chapter 4, let us now take a look at the specific details of the Event Detection process. The purpose of this chapter is to present the outputs generated from the Text Analysis Stage, together with how these is used to determine whether a sentence in an essay constitutes an Event and the output generated thereafter. The output generated from this stage is used as part of the score grouping process which makes up the next stage of the proposed solution.

In addition, a later section also details the performance of the Event Detection process, performed on a test bed of sample student essays, using performance measurements such as precision and recall and Matthew's correlation coefficient.

5.2 Events in a Narrative

As stated earlier in the previous chapters, Events can be seen to make up the core of a story. Therefore, it was hypothesized that by accurately detecting Events in an essay, it was possible to determine an essay's grade with respect to the criteria mentioned in Chapter 4.

Using Burke's (1969) previous work as a guide, an Event comprises:

- Actors
- Actions
- State

Hence, according to this, any sentence which contains these three instances would be considered an Event. Consider the following sentence:

"Daniel was walking over a bridge"

Here we can see that there is an Actor (Daniel), an Action (walking) and a Location (bridge) which is a State; therefore, we can classify this sentence as an Event. While this sentence may indeed be a rather simplistic one, it in fact introduces a character and establishes a 'bridge' as the setting; as such, it should be considered an important part of the story.

Events are separated from one another by noting the difference in States. Considering Table 5.1, in the first sentence, the character 'Daniel' is located on a bridge. However, in the next sentence he 'slipped', thereby landing in the river which is a different location.

Therefore, sentences 1 and 2 constitute separate Events. Of course it is naïve to assume that each Event has a different State, thus if this is not the case, those Events would then make up a Composite Event.

Taking Table 5.2 for example, in sentences 1 and 2 the character is in a cave. In sentence 3, his location changes from the cave to the hospital, thereby denoting a

separate Event. Therefore sentences 1 and 2 would make up a Composite Event, while sentences 3 and 4 would make up another.

1. Daniel was walking over a *bridge*.
2. Suddenly he slipped and fell into the *river*.

Table 5.1: Sample Sentences A

1. Daniel sat in a *cave*, his pants soaked from the wet, slippery floor.
2. He detested sitting down being immobilized by his ankle and decided to try and crawl around the *cave*.
3. Suddenly, he was lying on a *hospital* bed.
4. There were people all around him.

Table 5.2: Sample Sentences B

As such, the change in State is the main condition which determines whether or not an Event transitions from one to the other. However, realistically speaking, not every sentence can be considered an Event. Consider sentence 4 in Table 5.2. Although the sentence does not fulfil the three criteria for Event, it does not mean that it is entirely unimportant. Good stories are often made up of a few main events and other non-events that help establish the mood or tone of the story.

5.3 Text Analysis Output

The first step towards detecting Events is to process the text so that it is possible to extract, if any are available, the Named Entity Recognition (NER) tags of each word

followed by their POS tags. This stage is performed using POS and NER tools created by the Stanford Natural Language Processing Group, with some modifications to the NER tool to accommodate for the needs of this research project. The next two sections below will describe in detail the outputs generated by the above steps.

5.3.1 Named Entity Recognition

Even without customisation, the Stanford NER tool is still able to pick out most of the proper nouns without any additional training. Furthermore, for some purposes of this research project, the classifications provided by the NER tool (Location, Person, etc.) are also sufficient. As shown in the Figure 5.1, the tool is able to accurately pick out entities such as Persons and Locations.

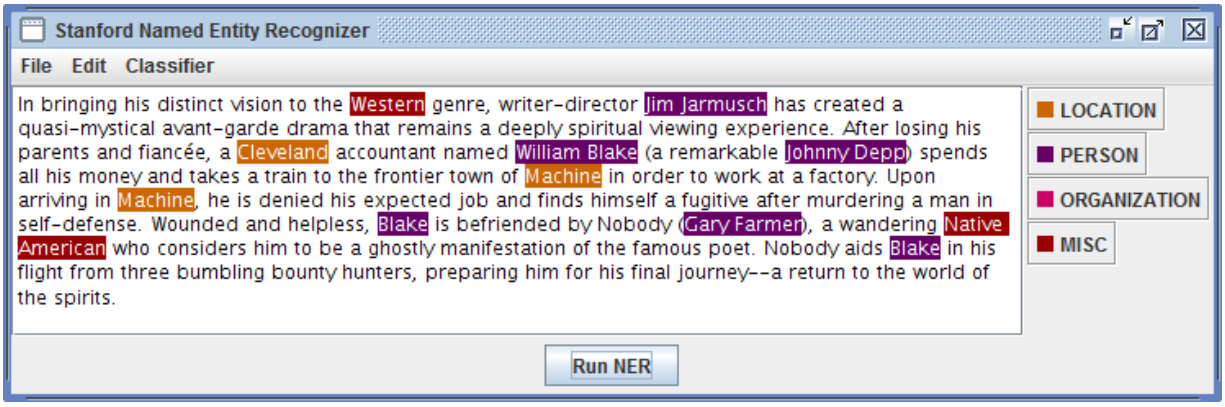

Figure 5.1: Output from Stanford NER tool

Note that the above examples are part of the Graphical User Interface (GUI) designed for end users of the NER tool. Since further analysis needs to be conducted in order to detect Events, the GUI is not used within the Event Detection framework.

Instead, for the purposes of this thesis, the NER classifiers are loaded onto a customised program written in JAVA and stored for further processing. Figure 5.2 shows the raw output:

Figure 5.2: Stanford NER tool raw output

For the full source code of the JAVA program, refer to Appendix D. After extracting the raw output, the next step is to extract the POS tag for each word.

5.3.2 Part of Speech Tags

Using the maximum entropy Part of Speech tagger, we are able to produce a POS tag for each word in the essay. Figure 5.3 shows the GUI which gives an example of words with their POS tags.

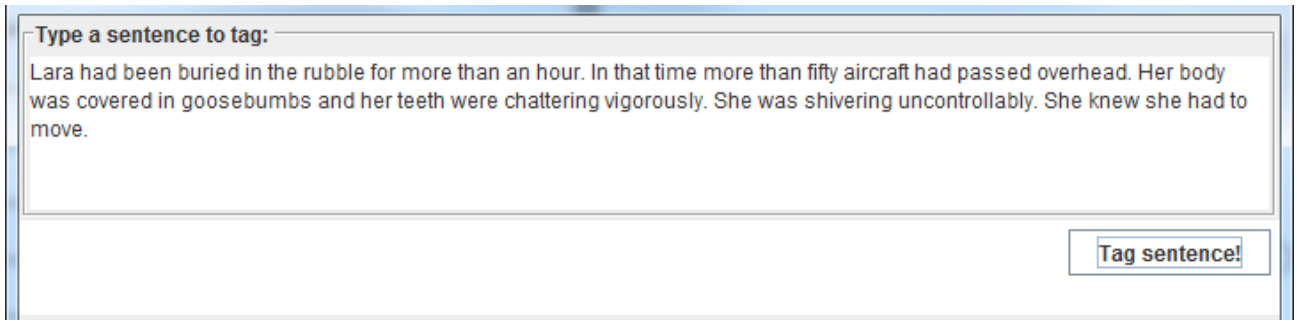

Figure 5.3: Stanford POS tagger output

Once again, for the purposes of Event Detection, only the raw output is needed as shown in Figure 5.4.

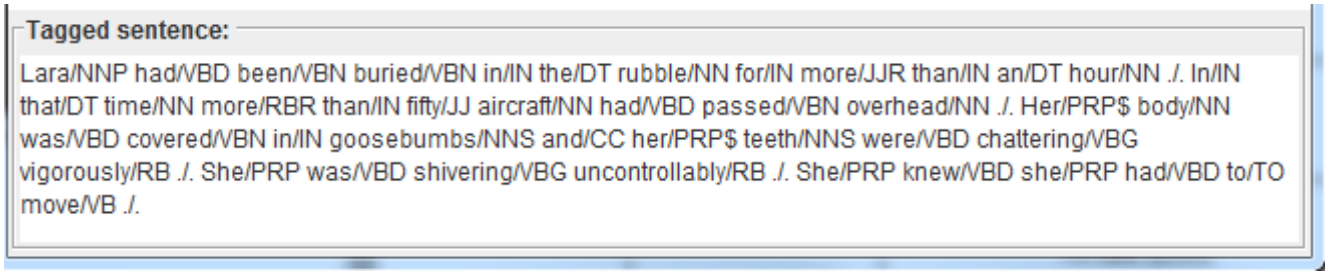

Figure 5.4: Stanford POS tagger raw output

Once both the NER and POS tags have been extracted, we are then able to perform the process of classifying a sentence according to whether or not it is an Event. This is done by combining the outputs generated and performing the Event Detection process.

5.4 Detecting Events

As mentioned earlier in Chapter 4, the NER tool had to be customised and trained to be able to pick out other representations of a State which is an essential part of an Event. Normally, the NER tool is able to pick out most types of Locations but narrative type essays require more 'fine tuning' to accurately pick out Locations. For example, Figure 5.5 shows that if only the default classifier is used, the tool picks out only one entity: Lara.

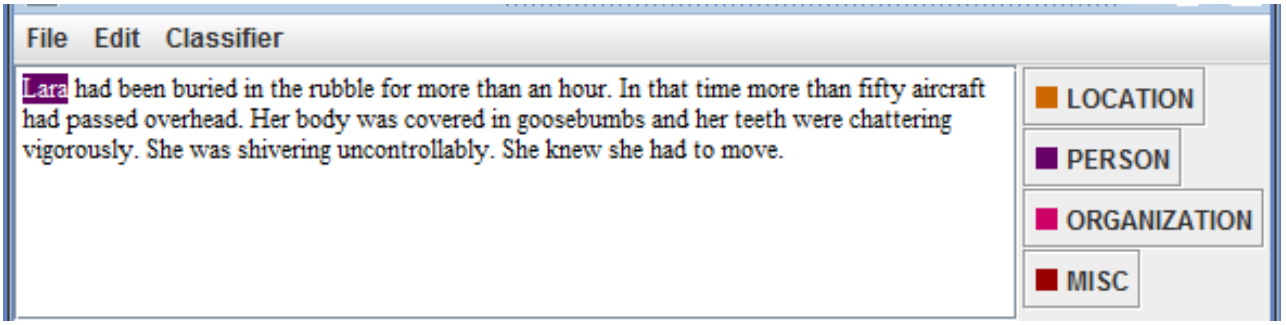

Figure 5.5: NER default classifier output

This output is not ideal for detecting Events since the phrase “in the rubble” should also be considered a Location. Using only the default classifier, this sentence would thus not fulfil the conditions to be classified as an Event, which is incorrect. Furthermore, the phrase “more than an hour” signifies the passage of time, which is also classified as a State.

Using the conditions of whether a word or group of words make up State (detailed in Chapter 4), the program goes through each sentence and if any are detected, the NER tag is replaced with the customised tag.

An example of the raw output of sentence 1 of the above sample is shown below:

```
Lara had been buried in the rubble for more than an hour.
=================================+
Lara  |     NNP   |     PERSON
had   |     VBD   |     O
been  |     VBN   |     O
buried|     VBN   |     O
in    |     IN    |     C.LOCATION
the   |     DT    |     C.LOCATION
rubble|     NN    |     C.LOCATION
for   |     IN    |     O
more  |     JJR   |     O
than  |     IN    |     STATE
an    |     DT    |     STATE
hour  |     NN    |     STATE
.     |     .     |     O
```

```
====================
```

From the output above, the program checks for an instance of each of the Event conditions, the result of which is shown in Table 5.3:

Actor	Lara
Action	Been buried
State	In the rubble More than an hour
Event	Yes

Table 5.3: Sample sentence 1 Event Classification result

With all three conditions fulfilled, the sentence would thus classify as an Event. The rest of the output generated from the essay sample excerpt is showed below, separated into individual sentences. This is followed by Tables 5.4 to 5.7 detailing each instance of an Actor, Action or State if detected, and the resulting Event classification.

Sentence 2

```
In that time more than fifty aircraft had passed overhead.
================================+
In    |     IN    |     STATE
that  |     DT    |     STATE
time  |     NN    |     STATE
more  |     RBR   |     O
than  |     IN    |     O
fifty |     CD    |     NUMBER
aircraft|   NN    |     O
had   |     VBD   |     O
passed|     VBN   |     O
overhead|   NN    |     O
.     |     .     |     O
====================
isEvent: NO
```

Actor	-

Action	Passed
State	In that time
Event	No

Table 5.4: Sample sentence 2 Event Classification result

Sentence 3

```
Her body was covered in goosebumbs and her teeth were chattering
vigorously.
================================+
Her   |     PRP$  |     O
body  |     NN    |     O
was   |     VBD   |     O
covered|    VBN   |     STATE
in    |     IN    |     STATE
goosebumbs |NNS   |     STATE
and   |     CC    |     O
her   |     PRP$  |     O
teeth |     NNS   |     O
were  |     VBD   |     O
chattering|VBG   |     O
vigorously|RB    |     O
.     |     .     |     O
====================
 isEvent: YES
```

Actor	Her
Action	Was, were
State	Covered in goosebumbs
Event	Yes

Table 5.5 Sample sentence 3 Event Classification result

Sentence 4

```
She was shivering uncontrollably.
=================================+
She   |          PRP   |     O
was   |          VBD   |     O
shivering|       VBG   |     O
uncontrollably|  RB    |     O
.     |           .    |     O
====================
 isEvent: NO
```

Actor	She
Action	Was, shivering
State	-
Event	No

Table 5.6: Sample sentence 4 Event Classification result

Sentence 5

```
She knew she had to move.
=================================+
She   |     PRP   |     O
knew  |     VBD   |     O
she   |     PRP   |     O
had   |     VBD   |     O
to    |     TO    |     O
move  |     VB    |     O
.     |     .     |     O
====================
 isEvent: NO
```

Actor	She
Action	Knew, had, move
State	-
Event	No

Table 5.7: Sample sentence 5 Event Classification result

5.5 Event Sequence and Ratio

The Event Sequence is represented by a set of 1s and 0s, which represent the sentence which is an Event and that which is not, respectively. The purpose of this sequence is to give a visualisation of the whole essay as a representation of Events and non-Events. Furthermore, it also allows us to view the distribution of Events over the entire length of the essay. This data would then later be used as an input when determining an essay's score according to the criteria mentioned in Chapter 4.

In addition to the Event Sequence, the output of this process also allows the Event Ratio to be displayed, which is calculated by dividing the number of Events by the total number of sentences in the essay. This provides another way of looking at the distribution of Events within the essay, which is similarly used in determining an essay's score with respect to certain criteria.

The resulting Event Sequence and Ratio for the above sample is shown in Table 5.8. The highlighted segment shows the first five sentences in relation to the Event sequence of the given sample.

```
EventCounter: 18 and NotEventCounter: 13
Event Sequence:
1 0 1 0 0 1 0 1 1 0 1 1 1 0 1 0 0 1 1 1 1 0 1 1 0 0 0 1 1 1
0
Event Ratio18/31=58.06%
```

Table 5.8: Event Sequence and Ratio

5.6 Testing and Evaluation

Since the ability to detect Events essentially forms the backbone of the proposed solution, it is imperative that this process be performed accurately. Towards this purpose, several tests were carried out to measure the performance of the Event Detection process.

The dataset used in this thesis is made up of 189 student essays ranging from Years 1 to 12, within the domain of narrative writing. The test bed was made up of 35 student essays that were above the mark of 25 out of a possible 47. The reason for this distribution was that since poorer essays usually have rather short scripts or just contain gibberish and few to no Events, they would not be an ideal test subject for determining the performance of this process.

An example of such an essay is shown below:

There was a Hipo that love to eta weeds that come in the water. But they only come once a year so the Hipo waited and waited up tell there was two more days to go the Hipo waited he was geting ready for

As stated earlier, short scripts such as the above sample would not be ideal for testing since they contain little content and are quite error prone. Therefore, for the specific purpose of measuring the performance of the Event Detection process, essays with a score below 20 were not used.

The abovementioned 35 essays were first evaluated by a human marker who manually annotated each sentence as an Event or non-Event. In total, there were 1340 sentences used in this part of the evaluation, with each essay averaging roughly 39 sentences.

Through human annotation, there were found to be a total of 682 instances where a sentence was classified as an Event and 675 instances where a sentence was a Non-Event. As shown in Figure 5.6, using the Event Detection process, a total of 658 Events were detected, with 665 sentences classified as a Non-Event. For details of the results regarding individual essays, refer to Appendix E.

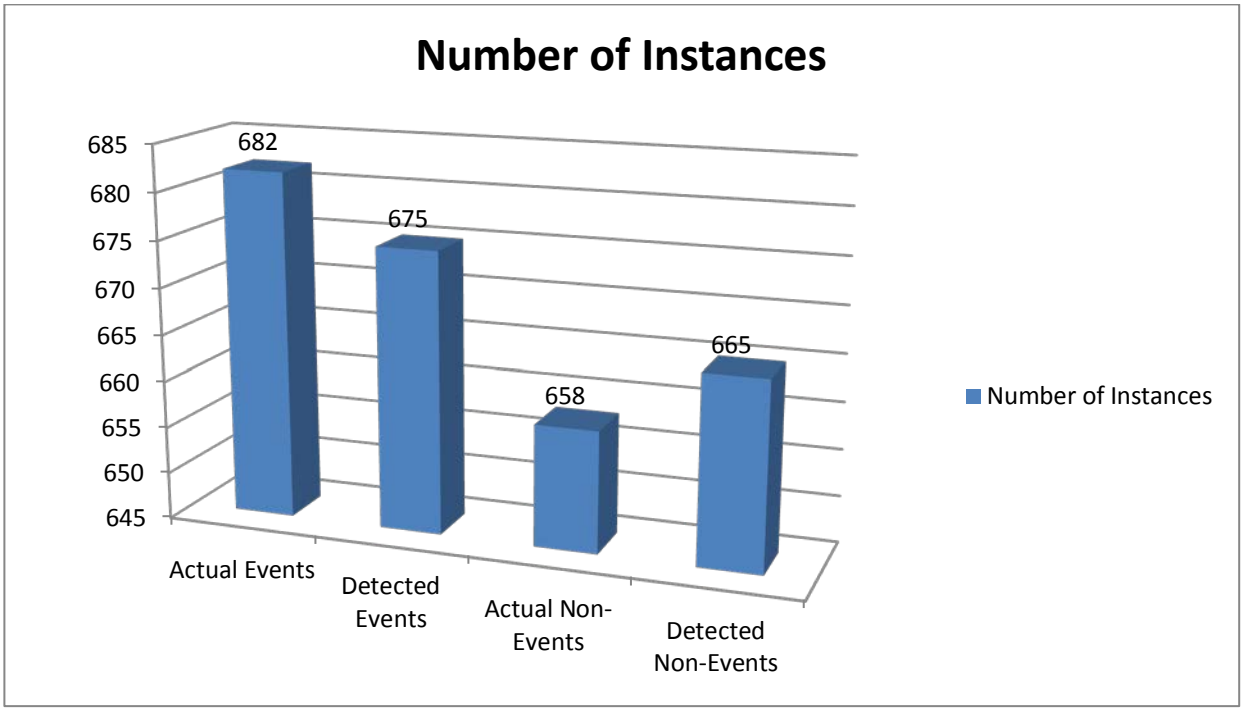

Figure 5.6: Event Classification Results

Looking at the overall results, one can assume that the agreement rate is relatively high (with a separation of only seven instances between both instances of Events and Non-Events) between the systems and a human annotator. However, the overall agreement

rate alone does not reflect cases of false positives, in which Non-Events are incorrectly classified as Events and false negatives, where Events are incorrectly classified as Non-Events.

Therefore, a more accurate measure of this process' performance would be to take into account the Precision, Recall and F-Measure scores.

5.6.1 Precision, Recall and F-Measure

Commonly, the evaluation of extraction techniques (summarization, information extraction) involves, but is not limited to, the use of two metric variables, namely Precision and Recall. Nenkova (2006) provides a simple definition of the two, stating that Precision is the number of instances in which the system was correct while Recall is the number of similar instances extracted by the computer and a human tester.

However, since the Event Detection stage is fundamentally a classification method, with regards to this thesis, Precision would thus refer to the number of sentences correctly classified as Events. Similarly, Recall would refer to the number of correctly classified sentences in relation to the total number of sentences that should be Events.

The formulas for calculating these are:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Formula 2: Precision

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Formula 3: Recall

Where:

- *TP = True Positives, the number of sentences correctly classified as Events*
- *FP = False Positives, the number of sentences incorrectly classified as Events*
- *FN = False Negatives, the number of sentences incorrectly classified as Non-Events*

Viewed in terms of a classification task, a score of 1 in Precision would thus mean that every sentence that the Event Detection process classifies as an Event is in fact an Event. Likewise, a score of 1 in Recall would mean that every sentence that should be an Event was correctly classified as such.

The result for the Precision and Recall for the Event Detection process is shown in Figure 5.7. For the full list of results stating the individual performance of the process on each essay, refer to Appendix G.

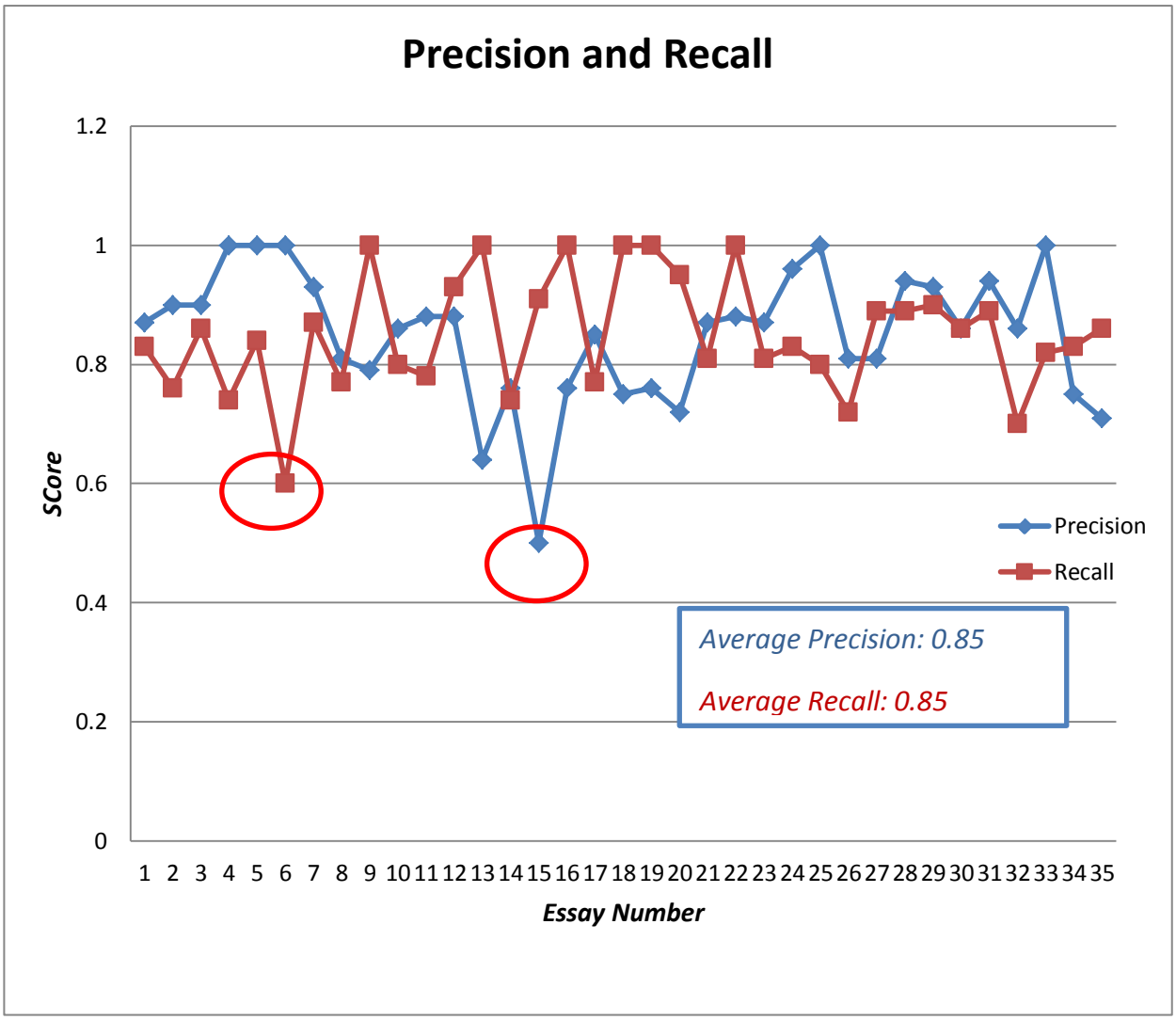

Figure 5.7: Precision and Recall Results

Judging from the results, the Event Detection process is rather promising, with an average of 0.85 in both Precision and Recall. The points highlighted with a red circle represent essays with a large disparity with the rest of the test set, which will be discussed in a later section.

Often, Precision and Recall are not taken as isolated measures and are instead considered together to measure a method's overall performance. One such value is the F-Measure, which considers both Precision and Recall together as a measure of a

system's accuracy. Hence, in determining the accuracy of the Event Detection process, the following formula is applied:

$$\text{F. Measure} = 2 * \left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\right)$$

Formula 4: F-Measure

Once again, a score of 1 indicates a 100% classification success rate. Therefore, the objective is to attain an F-Measure as close to 1 as possible. Figure 5.8 below shows the F-Measure for each essay, together with the average score.

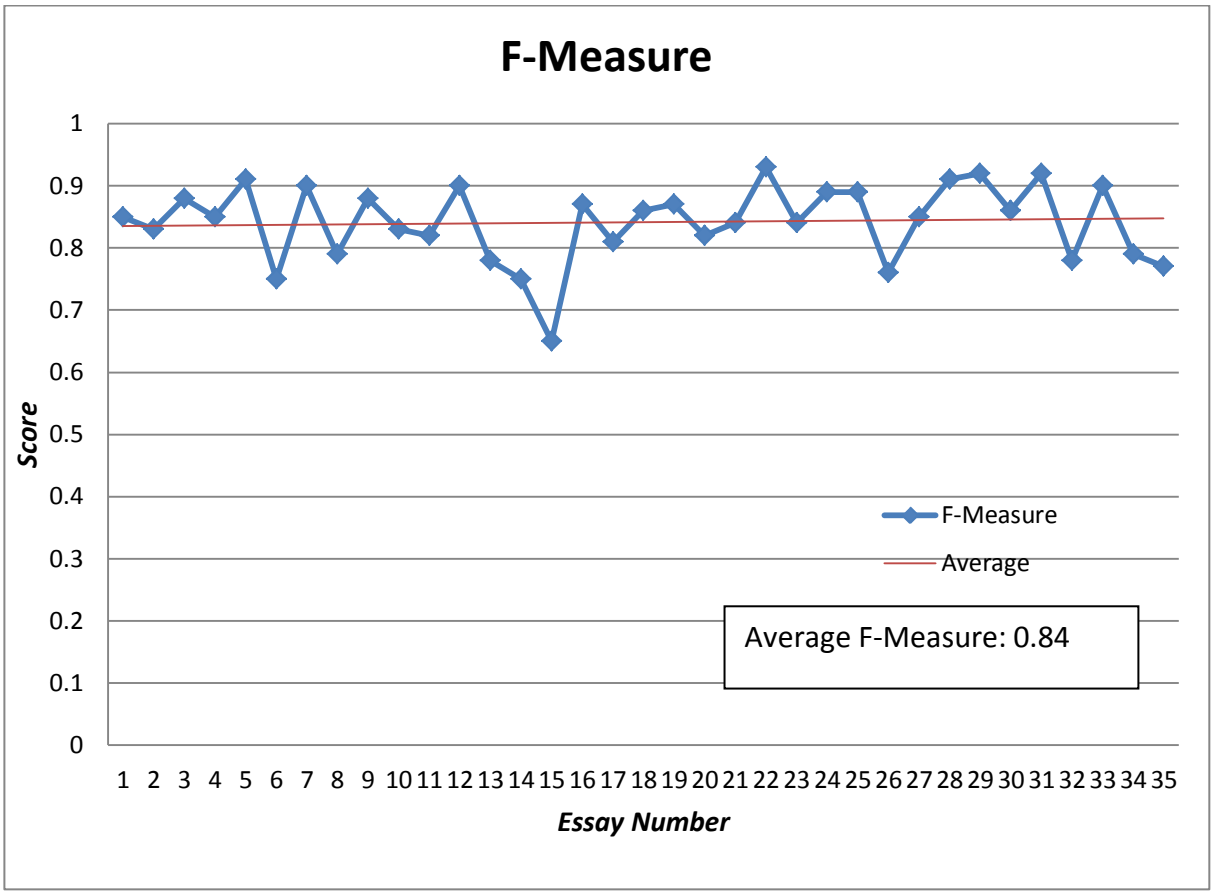

Figure 5.8: F-Measure Results

To give a better indication of the performance of this process, the total number of true positives and negatives, together with the number of false negatives, were taken into account in order to provide an overall measure of its performances. Through manual annotation of the results, the total number of the abovementioned variables were calculated and applied to the Precision, Recall and F-Measure algorithms. The results are shown in Figure 5.9 below:

True Positives	True Negatives	False Positives	False Negatives	Total
568	556	103	113	1340

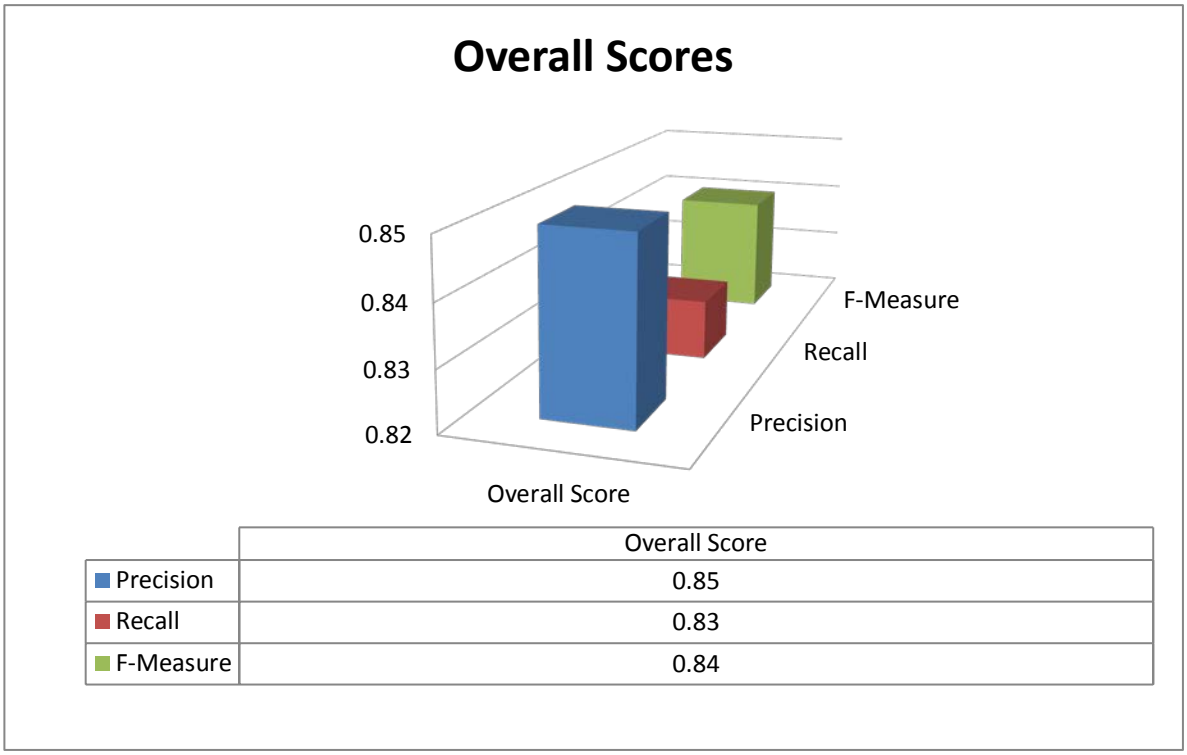

Figure 5.9: Overall scores for Precision, Recall and F-Measure

Thus far, the process has shown rather encouraging results, with an average of 0.85 in Precision and Recall and an average F-Measure of 0.84. In general, the performance

varies little, with an overall Precision, Recall and F-Measure scores of 0.85, 0.83 and 0.84 respectively.

There is however, one more issue that needs to be addressed. The Event Detection process should be considered as a binary classifier since it determines which sentences are Events and which are not, as opposed to others in which the true positives are the main focus. Therefore, while the F-Measure is a good indicator of the combined equal weighting of Precision and Recall, it does not take into account the rate of true negatives (TN). This is a problem particularly when there are a high number of true negatives but few true positives, which would give a lower Recall score and hence a lower overall F-Measure.

5.6.1.1 Addressing Errors

With regards to Precision, essay 15 is an outlier with a score of only 0.5. A lower score in Precision indicates a high number of false positives, where the classifier mistakenly classifies a sentence as an Event when in fact it is not. Upon closer inspection of essay 15, the reason for the high number of false positives is the higher number of errors in the pattern recognition step of the State Detection stage. Although this would be a potential problem in other essays of a similar nature, the impact on the overall essay grading process might be less due to the fact that all 35 essays had similar grades. This means that occurrences of essays of this type are fewer and are easily picked out for closer examination by a human marker.

Addressing issues in Recall, essay 6 receives a score of only 0.6, which might mean that there were a large number of false negatives. Upon closer inspection however, it was found that this was in fact due to the large number of true negatives, because the process has correctly identified which sentences were not Events. This value is not taken into account when measuring Recall, thereby increasing the value of the denominator in the algorithm while the numerator remains low due to a low number of true positives, resulting in a lower Recall score. The total number of errors found in essay 6 was in fact only 5, out of a total of 58 sentences, with 0 false positives, thus giving it a perfect Precision score of 1.

In order to more effectively determine the performance of the process, the performance measurement of Matthew's Correlation Coefficient (MCC) was used.

5.6.2 Matthew's Correlation Coefficient

This measure was first used by Matthews (1975) and is closely related to Pearson's correlation coefficient, albeit in the context of secondary structure predictions. One of the main reasons that this measure is used to evaluate the performance of this process is because it allows us to take into account the true negative rate. Another advantage of this measure is that it still allows for a fair assessment even if the test classes are of very different sizes. The formula for Matthews Correlation Coefficient is shown below:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Formula 5: Matthew's Correlation Coefficient

Where:

- *TP = True Positives*
- *TN = True Negatives*
- *FP = False Positives*
- *FN = False Negativee*

The return value from the above algorithm ranges from +1 to -1, with a value of 0 indicating the performance of a random classifier. An MCC value of 1 would mean that the classifier performs perfectly while a value of -1 indicates total disagreement with the desired result. Therefore, similar to the previously used performance measures, a value as close to 1 as possible is the objective. Figure 5.10 shows the MCC value for each essay according to its individual TP, TN, FP and FN values.

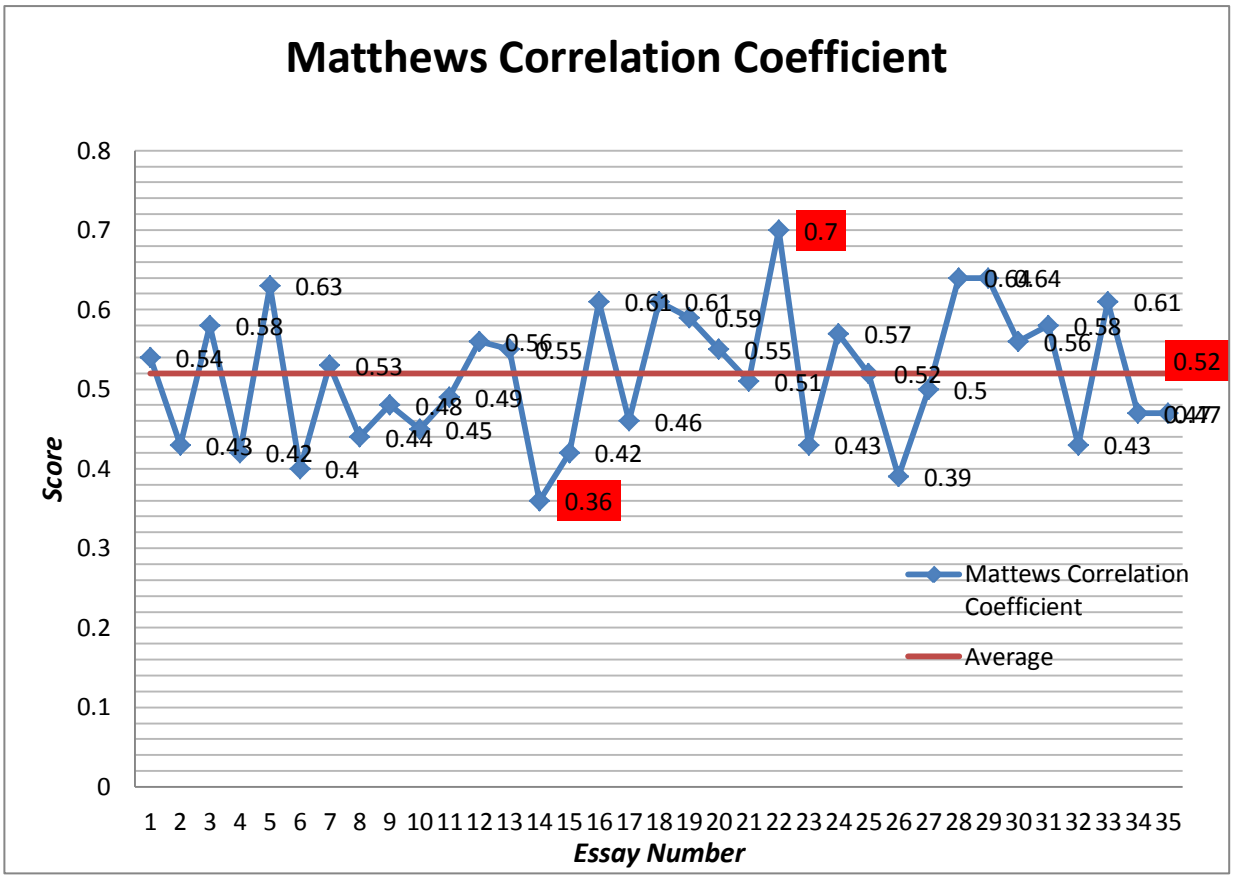

Figure 5.10: Matthew's Correlation Coefficient result

From the results, it can be seen that the MCC value ranges from 0.36 at its lowest and 0.7 at its highest, with an average value of 0.52.

As with the previous performance measures, to give a better indication of the overall performance with regards to Matthews Correlation Coefficient, the total TP, TN, FP and FN values were taken into account; the resulting value is the same as the average at 0.52.

5.7 Conclusion

This chapter presented in detail the steps involved in the Event Detection process. Each essay is split up into its individual sentences and classified as an Event or Non-Event according to the Event criteria.

As mentioned earlier, this stage forms the foundation on which the proposed solution is based, making it imperative that it perform at an acceptable level. This objective appears to have been achieved with overall Precision, Recall and F-Measure scores averaging over 0.80, a rather promising score for an otherwise untested method.

Using the Matthews Correlation Coefficient as a more effective performance measurement showed more detail in analysing the success rate of the classifier. With scores ranging from 0.36 to 0.70 and an overall of 0.52, it can definitely be said that classifier does not perform at a seemingly random success rate.

However, issues such as the high number of errors in one essay due to the inaccuracy of the pattern recognition step are cause for concern. Although it might have little impact on the overall success rate of the proposed solution, it is by no means a problem to be overlooked. Possible solutions include further training for the State detection method and more in-depth contextual analysis to improve accuracy.

Overall, the Event Detection process performs at a reasonably acceptable level for its results to be used in the next stage of the proposed solution. The next chapter will discuss how the output gathered from this process is applied to determine an essay's score group.

Chapter 6 - Group Scoring for Audience

6.1 Introduction

The previous chapters have described the components of an Event and how these are detected. In this chapter, the focus is on the audience criterion and how an essay is automatically scored based on the formalisation of the NAPLAN marking rubric. To briefly reiterate what was discussed in the previous chapter, an Event is made up of three main components, namely:

- Actor
- Action
- State

Using a Part of Speech tagger together with a customised Named Entity Recognition tool, each sentence in an essay is scanned to see if it contains one or more of the aforementioned components. If those components are present, then that sentence is classified as an Event. Apart from determining whether or not a sentence can be classified as an Event, several other details of an essay are acquired by the Event Detection process. These include:

- Essay length
- Event Ratio
- Number of Events

- Presence of physical or mental States
- Number of words used
- Number of unique and total adjectives
- Number of unique and total verbs
- Number of unique and total nouns

These are just some of the details that can be gathered by the Event Detection process discussed in Chapter 5, although not necessarily all of these variables are utilised when conducting experiments for each criterion.

6.1.1 Overview of the Rubric Formalisation Process

With the Event Detection output generated, the next step is to apply the previously discussed scoring logics to the acquired data. According to the NAPLAN rubric, the audience criterion was made up of six individual band scores as shown below:

Band	Description
0	Symbols or drawings which have the intention of conveying meaning
1	Contains some written content
2	Shows awareness of basic audience expectations through the use of simple narrative markers
3	An internally consistent story that attempts to support the reader by developing a shared understanding of context
4	Supports reader understanding and attempts to engage reader

5	Supports and engages reader through deliberate choice of language and use of narrative devices
6	Caters to the anticipated values and expectations of the reader Influences or affects the reader through precise and sustained choice of language and use of narrative devices

These were then manually divided into three main Score Groups: Poor, Intermediate and Good. There are two reasons for this: firstly, the difference between the higher band scores (5 and 6) is extremely subtle and subjective; hence, it is crucial to provide a simpler representation of the rubric. It was assumed that on a larger point scale (the total achievable mark was 47) a difference of 1 point would have little effect on the final overall grade. The validity of this assumption is discussed in Chapter 10.

Secondly, it allows students to receive feedback specific to the criterion itself. This is valuable since the overall grade, be it high or low, does not provide a clear indication of strengths or of areas that need improvement.

Thus, the Audience criterion is separated into the three groups as follows:

	Poor	*Intermediate*	*Good*
Audience	1-3	4	5-6

The features considered when allocating an essay to its appropriate Score Group for Audience are:

- Essay length
- Number of Events
- Event Ratio
- Physical and/or Mental State

The details according to which the essays are placed in the 0 Score Group are not discussed here since a separate project carried out by another researcher will filter out these essays before they are processed for grading, placing them outside the scope of this thesis.

As mentioned in Chapter 4, a two-step process is used to determine the score group of an essay. Firstly, the features which meet the Audience criterion for a particular band score need to be determined. Once done, these features are weighed according to their significance in relation to this criterion.

The second step takes into account the particular specified conditions previously determined that place an essay in one of the three categories namely A, B or C in relation to the Audience criterion. Once these steps have been performed, the output from both these processes is combined and used to determine the audience score group to which an essay belongs. Taken from Chapter 4, Figures 6.1 and 6.2 below illustrate the Grouping and Score Grouping processes respectively.

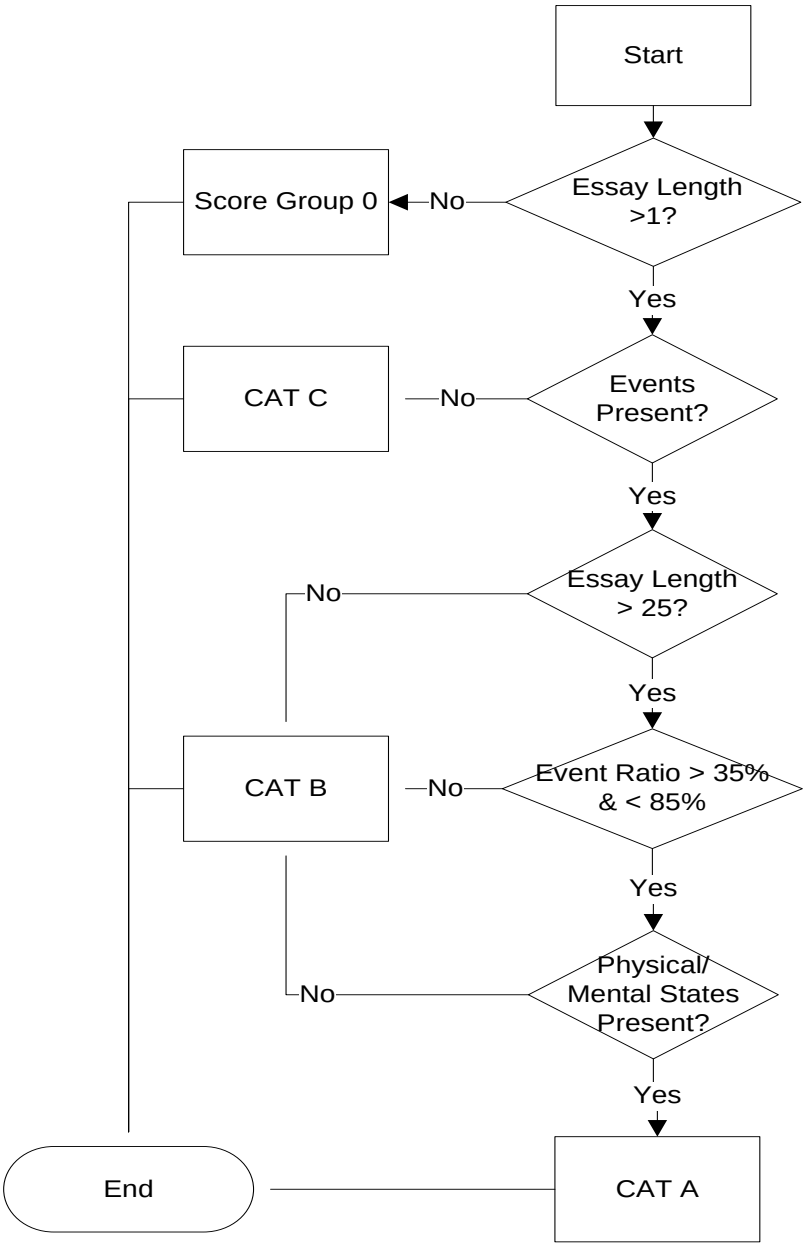

Figure 6.1: Grouping for Audience, from Chapter 4

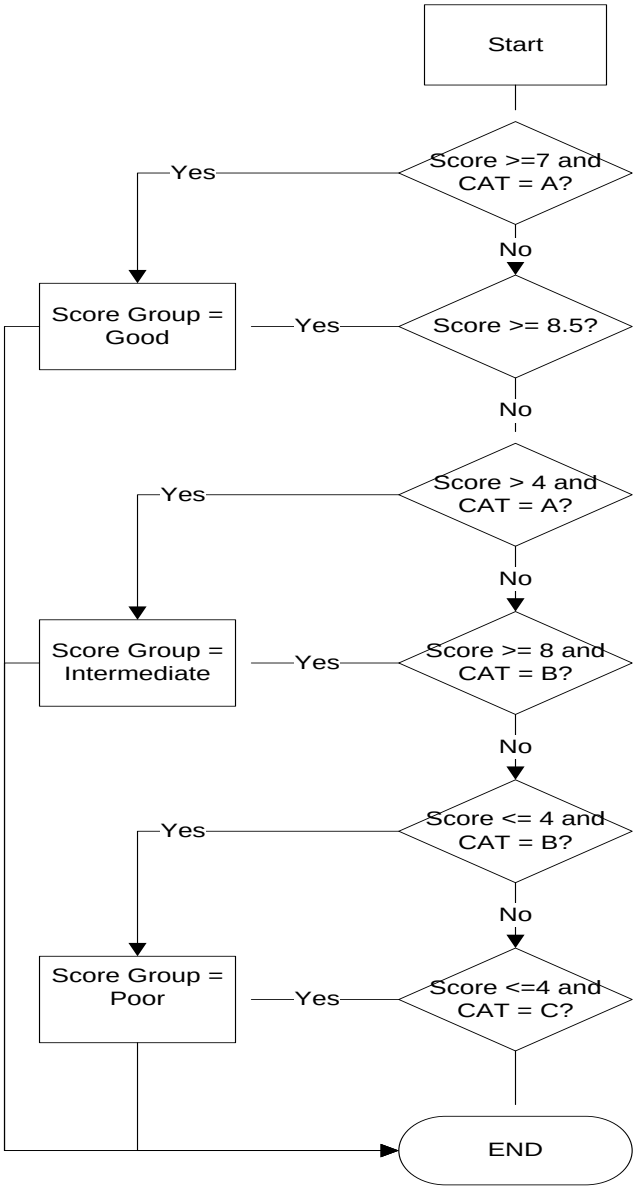

Figure 6.2: Score grouping for Audience, from Chapter 4

The main aim of this chapter is to ensure that the methods used to Score Group the essays are in fact valid. Hence, the rest of this chapter will detail the methods used to carry out this experiment and will state the hypothesis which tests the validity of the methods. The chapter then concludes with an analysis of the results gathered from these experiments.

6.2 Methodology Stage 1 – Precision, Recall, F-Measure and Exact Agreement

One hundred and eighty-nine student essays were used in this experiment, out of which 99 were allocated to training the scoring logics. The evaluation was performed in two main stages; the first compared the system with other current automated grading systems using Precision, Recall, F-measure and Exact Agreement values.

The first stage consisted of two steps. Firstly, the performance metrics of Precision, Recall and F-measure in which the human marker assigned scores were used as the basis of evaluation were calculated. The second step established the exact agreement rate between the number of essays placed in a particular Score Group by the human markers and those allocated to the same group by the system (machine).

The second stage of the evaluation tested the hypothesis using the Pearson's Chi Squared and the paired T-test statistic.

6.2.1 Pre-Experiment Details

During the first stage, each essay was manually sorted into its specific Score Group of Poor, Intermediate or Good, according to the band score it received for that particular criterion with relation to the NAPLAN marking guide.

The scoring logic was then calibrated such that the exact agreement rate was at an acceptable level before the testing phase was carried out. The other half, consisting of 90 essays was then used to test the system's accuracy. Similar to the training phase, each essay was first placed by human markers in its relevant Score Group in relation to

the band score it received, as Table 6.1 shows. For the full list of band scores assigned to all essays, please refer to Appendix F.

	Audience Score Group		
	Poor	*Intermediate*	*Good*
Essay Name	ADKINS	ADANO	ANDREWS
	AGENBAG	AMESS	BAGIATIS
	BEAVEN	AZMI	BAKER.C
	BENNETT	BAKER.L	BELLIS
	BERTOLA	BERENTE	BOCCAMAZZO.C
	BRIGGS	BETTI	BOLES-RYAN
	CHEREL	BIRSS	BOTHMA
	CHETWYND	BOCCAMAZZO.D	BOTH-WATSON
	COMBI	BRAMPTON	BOWEN
	COYNE	CHARLES	BREAN
	DALE-FRASER	CHU	BYRNES
	DARCEY	COLBY	CABUNALDA
	DOPOE	CONN	CASTAING
	ELLERTON	CUNNINGHAM	CATOVIC
	FARLEY	DE PLEDGE	CHANDLER
	FARRELL	DIXON	CHEDID
	FERGUSON	DORRELL	COPPARD
	GUTHRIE	ESTENS	COWELL
	HAGUE	GALANTE	DANKS
	HALL	GIANATTI	DORAN
	HANSEN.TA	GORJY	FOO
	HANSEN.TR	HAINES	FORWARD
	HARLAND	HELSBY	GESTE
	HENRY	HIGGINSON	HAWKETT
	HODSON	HOLT	KROLL
	HUDSON	INGRAM	LOH
	HUNTER	JOHNSON	MAIN
	IOPPOLO	KARSKI	PALAYUKAN
	JONES		PERSSON
	KELLY		VICKERY
	MASON		
	MILTON		

Table 6.1 Individual Essay Score Groups according to Human Markers - Audience

Similar to the Event Detection stage, a perfect score in Precision in this instance would mean that every essay the system classified as belonging to a particular Score Group did indeed belong there, while a perfect score in Recall states that every essay that was meant to be in that particular Score Group was classified as such. Using the F-measure

as a combination of the aforementioned measures, the system accuracy was then determined.

6.2.2 Exact Agreement Rate

The performance measure used in the second step of this stage of the evaluation was the exact agreement rate between the human markers and the system. Since the essays were previously sorted into their respective Score Groups according to the scores assigned by the human markers, the rate at which the system groups an essay within the same Score Group can be taken to be the exact agreement rate. According to Larkey (1998), the exact agreement rate is ideal for capturing the degree of similarity between one scoring procedure and another.

In this measure, the number of instances where the human markers and system both placed an essay within the same Score Group is determined, which thus allowed us to compare the extent to which the system's assessment of an essay is similar to the assessment by its human counterparts. The exact agreement rate for each of the scoring logic is calculated using the following formula:

$$\boldsymbol{Exact\ Agreement\ Rate} = \frac{\mathbf{NS}}{\boldsymbol{NH}}$$

Formula 6: Exact Agreement Rate

Where:

- NH = the number of essays placed in a Score Group according to human marker assigned scores
- NS = the number of essays placed within the same score group by the system

6.2.3 Experiments and Results

As with the Event Detection process, the objective is to achieve as close to a score of 1 as possible in Precision, Recall and F-measure values. However, a perfect score in all areas is neither a feasible nor realistic benchmark to go by since even among human markers the correlation is rarely perfect; the same can be said for the agreement rate. From the literature review conducted in Chapter 2, it was found that the average accuracy between systems that used such a performance measure was 0.91, whereas the only system that used the exact agreement rate showed a value of 0.55.

Since the F-measure is a combination of the Precision and Recall values, it can be taken to represent the accuracy of the system. Therefore, for the purposes of this thesis, the objective is to achieve an F-measure score as close to 0.91 as possible while faring no less than 0.65. In addition, the average values for Precision, Recall and F- Measure across all Score Groups should be at least above 0.65. In terms of the exact agreement rate, the objective is to achieve an overall average as close to 100% as possible while faring no less than 55%.

The next sections discuss the experiments carried out within each Score Group and the subsequent results, followed by the testing of the hypothesis.

6.2.3.1 Score Group - Poor

According to band scores assigned by human markers, 32 essays were found to be in this Score Group. Under the scoring logic, the system classified 34 essays as belonging to this particular group.

Table 6.2 shows the results where the first column, Actual, lists the essays that should be within this Score Group while the second shows the list of essays correctly classified by the scoring logic. The third and fourth columns show the essays that were erroneously placed in or outside of this Score Group respectively.

Actual	System (True Positives)	Errors	
		False Positives	False Negatives
ADKINS	ADKINS	AZMI	NIL
AGENBAG	AGENBAG	COLBY	
BEAVEN	BEAVEN		
BENNETT	BENNETT		
BERTOLA	BERTOLA		
BRIGGS	BRIGGS		
CHEREL	CHERAL		
CHETWYND	CHEYWYND		
COMBI	COMBI		
COYNE	COYNE		
DALE-FRASER	DALE-FRASER		
DARCEY	DARCEY		
DOPOE	DOPOE		
ELLERTON	ELLERTON		
FARLEY	FARLEY		
FARRELL	FARREL		
FERGUSON	FERGUSON		
GUTHRIE	GUTHRIE		
HAGUE	HAGUE		
HALL	HALL		
HANSEN.TA	HANSEN.TA		
HANSEN.TR	HANSEN.TR		
HARLAND	HARLAND		
HENRY	HENRY		
HODSON	HODSON		
HUDSON	HUDSON		
HUNTER	HUNTER		
IOPPOLO	IOPPOLO		
JONES	JONES		
KELLY	KELLY		
MASON	MASON		
MILTON	MILTON		

Table 6.2: Audience Score Grouping results – Score Group "Poor"

6.2.3.1.1 Discussion

From the values gathered, the scoring logic shows promising results, with a Recall value of 1 indicating that every essay that should be in the Score Group "Poor" was indeed classified as such. All 32 essays that were supposed to be within this group were detected while the remaining 2 essays were found to belong to the next better Score Group, Intermediate.

Based on the results above, the Precision, Recall and F-Measure values were calculated, the results of which are shown in Table 6.3.

True Positives	False Positives	False Negatives	Precision	Recall	F - Measure
32	2	0	0.94	1	0.97

Table 6.3: Precision, Recall and F-Measure results for Audience – Score Group "Poor"

False positives indicate that the system might be slightly more stringent than human markers, with 2 essays being placed in a lower group than they should be. A high F-measure value of 0.97 indicates a high accuracy when identifying essays within this Score Group.

The exact agreement rate for this scoring logic came to 100%, in that for every essay that the human markers placed in this Score Group, the system did as well, as shown in Table 6.4. Even though there were 2 additional essays that the system placed here, this is considered an error in the scoring logic for the "Intermediate" Score Group and is thus addressed in the respective section.

Score Group	System	Human Makers	Exact Agreement Rate
Poor	32	32	100%

Table 6.4: Exact Agreement Rate for Audience – Score Group "Poor"

6.2.3.2 Score Group – Intermediate

According to Table 6.5, 28 essays were deemed to belong to this Score Group with a total number of 14 essays detected by the system. The number of false negatives was significantly higher than the scoring logic for the previous Score Group, with 14 essays being classified as belonging to a different group.

The Recall value suffered from the high number of false negatives, where the system failed to correctly place Intermediate essays within the correct Score Group. Of the 14 false negatives, 2 were found in the "Poor" Score Group while 12 were placed in the Score Group "Good". The 8 essays that were placed in this group were scored higher by human markers.

Actual		System (True Positives)	Errors	
			False Positives	False Negatives
ADANO AMESS AZMI BAKER.L BERENTE BETTI BIRSS BOCCAMAZZO.D BRAMPTON CHARLES CHU COLBY CONN CUNNINGHAM DE PLEDGE DIXON DORRELL	ESTENS GALANTE GIANATTI GORJY HAINES HELSBY HIGGINSON HOLT INGRAM JOHNSON KARSKI	BETTI BOCCAMAZZO.D CHARLES CHU CONN CUNNINGHAM DORRELL ESTENS GALANTE GORJY HIGGINSON INGRAM JOHNSON KARSKI	BAKER.C BELLIS BOWEN BREAN BYRNES CATOVIC CHANDLER FORWARD	ADANO AMESS AZMI BAKER.L BERENTE BIRSS BRAMPTON COLBY DE PLEDGE DIXON GIANATTI HAINES HELSBY HOLT

Table 6.5: Audience Score Grouping results – Score Group "Intermediate"

6.2.3.2.1 Discussion

This section of the scoring logic returned poorer results than expected, with Precision and Recall values at 0.64 and 0.50 respectively, shown in Table 6.6. The distribution of false negatives was also evenly spread between the other two Score Groups namely "Poor" and "Good", meaning that the system is neither more lenient nor stricter than its human counterparts. This might mean that further calibration of the scoring logic is required. Other solutions might be to take into account other features of the text as well as incorporating more in-depth contextual analysis.

True Positives	False Positives	False Negatives	Precision	Recall	F - Measure
14	8	14	0.64	0.50	0.56

Table 6.6: Precision, Recall and F-Measure results for Audience – Score Group "Intermediate"

With the total number of essays placed in a different Score Group, the agreement rate for this scoring logic came to 50%. However, the system agreed only with the human markers on 14 essays out of the 28 placed in this group according to the human marker assigned scores. Table 6.7 elaborates.

Score Group	System	Human Makers	Exact Agreement Rate
Intermediate	14	28	50.00%

Table 6.7: Exact Agreement Rate for Audience – Score Group "Intermediate"

6.2.3.3 Score Group – Good

This last group had a total of 30 essays as determined by the band scores assigned by human markers. Through the scoring logic, the system correctly identified 22 essays

that belonged to this group with 8 essays scoring lower and 12 being higher than the human marker scores. Table 6.8 elaborates.

Actual		System (True Positives)	Errors	
			False Positives	False Negatives
ANDREWS	COWELL	ANDREWS	ADANO	BAKER.C
BAGIATIS	DANKS	BAGIATIS	AMESS	BELLIS
BAKER.C	DORAN	BOCCAMAZZO.C	BAKER.L	BOWEN
BELLIS	FOO	BOLES-RYAN	BERENTE	BREAN
BOCCAMAZZO.C	FORWARD	BOTH-WATSON	BIRSS	BYRNES
BOLES-RYAN	GESTE	BOTHMA	BRAMPTON	CATOVIC
BOTHMA	HAWKETT	CABUNALDA	DE PLEDGE	CHANDLER
BOTH-WATSON	KROLL	CASTAING	DIXON	FORWARD
BOWEN	LOH	CHEDID	GIANATTI	
BREAN	MAIN	COPPARD	HAINES	
BYRNES	PALAYUKAN	COWELL	HELSBY	
CABUNALDA	PERSSON	DANKS	HOLT	
CASTAING	VICKERY	DORAN		
CATOVIC		FOO		
CHANDLER		GESTE		
CHEDID		HAWKETT		
COPPARD		KROLL		
		LOH		
		MAIN		
		PALAYUKAN		
		PERSSON		
		VICKERY		

Table 6.8: Audience Score Grouping results – Score Group "Good"

6.2.3.3.1 Discussion

The results here show a slight improvement in Recall and F-measure scores, with values of 0.69 and 0.64 respectively. The lower Precision value is attributed to the higher number of false positives when compared to the number of correctly classified essays. This leads to the assumption that the scoring logic is slightly more lenient in its assessment.

Based on the results gathered from this section of the scoring logic, the subsequent Precision, Recall and F-measure values are as shown in Table 6.9.

True Positives	False Positives	False Negatives	Precision	Recall	F - Measure
22	12	8	0.60	0.69	0.64

Table 6.9: Precision, Recall and F-Measure results for Audience – Score Group "Good"

With the system and its human counterparts agreeing on 22 out of the 30 essays found to belong to this Score Group, the exact agreement rate came to 73%, as shown in Table 6.10.

Score Group	System	Human Markers	Exact Agreement Rate
Good	22	30	73.33%

Table 6.10: Exact Agreement Rate for Audience – Score Group "Good"

6.3 Methodology Stage 2 -Hypotheses Testing

There are two main tests used in this work when testing the hypothesis. The first is a goodness-of-fit test, also known as the Pearson's Chi-Squared Goodness of Fit test, which tests whether or not the data observed is of a random nature. The second test is the paired T-test, which tests for any significant difference between the mean score of the paired difference between two sample groups and a specified mean value.

6.3.1 Hypotheses

Based on the features described in the conceptual framework of group scoring essays under the Audience criterion, the following hypotheses were generated for this part of the thesis:

The null hypothesis $\mathbf{H_0}$ is formulated as:

> $\mathbf{H_0}$ **:** There is no significant difference between the human marker scores and the machine-generated scores for the Audience criterion.

The alternate hypothesis $\mathbf{H_1}$ would thus be:

> $\mathbf{H_1}$: There is a significant difference between the human marker scores and the machine generated scores for the Audience criterion.

The sections below give a brief overview of the Chi-squared and paired T-test.

6.3.2 Chi-Squared Goodness of Fit Test

As mentioned earlier, this test is used to determine the 'goodness of fit' of the data used. In other words, it shows how close the values of the observed data are to those of the expected values (Plackett 1983). The formula is:

$$x^2 = \sum \frac{(O - E)^2}{E}$$

Formula 7: Chi Squared Goodness of Fit test

Where:

- O = observed data in each category

- E = Expected value

Once obtained, the Chi-square value can then be used to obtain the probabilities (P values) from a Chi-square distribution Table. The P value allows us to determine whether the observed deviations are due to random chance alone according to the degrees of freedom, which is the number of categories minus 1.

6.3.3 Paired T-Test

The simple T-test allows us to assess whether the means of two groups are statistically different from each other. This allows us to determine whether the difference between those means is significantly more or less than zero. Once the t value is acquired, it is compared against a Table of critical t values which determines whether or not the difference is significant (Skoog 2003).

The formula for conducting a simple t test is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

Formula 8: Simple T-test

Where:

- $\overline{x}$ = the mean of a sample group
- S = the variance of a sample group
- n = the sample size

However, since the values of means from the two data groups are related, the paired T-test is used instead. The formula for this is:

$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{n}}.$$

Formula 9: Paired test

Where:

- $\overline{x}$ = the sum of the difference between the means
- μ = the expected mean
- SD = standard deviation of X
- N = sample size

In terms of this thesis, it is assumed that the human markers' scores and those generated by machine should be similar. Therefore, the expected mean value between the differences of the two scores should be as close to zero as possible, allowing us to conclude that there is no significant difference between the two.

6.3.4 Chi-Squared Goodness of Fit Test Results

In conducting the goodness of fit test, the number of essays classified as poor, intermediate or good were counted and sorted according to machine and human marker scores. For the purposes of the Chi-square test, the machine provides the observed values while the human marker scores are the expected values with a degree of freedom of 2, the values of which are shown below:

Group	Observed	Expected
Poor	34	32
Intermediate	22	28
Good	34	30

Based on the data, the formula for determining the Chi-square was:

$$x^2 = \frac{(34-32)^2}{32} + \frac{(22-28)^2}{28} + \frac{(34-30)^2}{30}$$

$$x^2 = 0.125 + 1.28 + 0.53$$

$$x^2 = 1.938$$

DF	P=0.995	P=0.975	P=0.9	P=0.5	P=0.1	P=0.05	P=0.05	P=0.01	P=0.005
1	0.000	0.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838

Table 6.11: Chi Square Distribution Table

In order to reject the null hypothesis, the value of P should be greater than 0.95. When compared to the highlighted row (DF 2) on the Chi-square distribution Table (refer to Appendix H for full Table), the P value obtained is between 0.5 and 0.10, which is insufficient to reject the null hypothesis.

6.3.5 Paired T-test Results

If the null hypothesis is to be rejected, the t value needs to be significant when compared against the Table of critical *t* values. The resultant formula when conducting the test was:

$$t = \frac{0.03 - 0}{\sqrt{0.52/(89)}}$$

$$t = 0.52$$

DF	0.1	0.05	0.02	0.01	0.005	0.002	0.001
89	1.6622	1.9870	2.3690	2.6322	2.8787	3.1844	3.4032

Table 6.12: Critical T values Table at 89 degrees of freedom

In order for there to be a significant difference between the machine and human marker scores, the value of t should exceed the t critical value which is at 1.98. As shown in the Table above, the t value obtained is much lower than the critical value, at 0.52.

Based on the resulting evidence, the null hypothesis $\mathbf{H_0}$ cannot be rejected; thereby leading to the conclusion that there is insufficient evidence to suggest that there is a significant difference at a 95% confidence level between the scores generated by the machine and those arrived at by human markers.

6.4 Conclusion

This experiment was conducted using 90 essays selected from a group of 189 essays, of which 99 were used to calibrate the system. Overall, the system achieved an agreement rate of exactly 73% with its human counterparts, with a total of 24 erroneously classified essays. Of these 24, 10 essays were scored lower, while 14 were placed in a higher Score Group by the system.

Recall and F-measure values showed good results, achieving an average score ranging from 0.72 to 0.73 across all performance measures used in this experiment. This enables us to conclude that the features identified earlier relate well to the score an essay receives against the Audience criterion.

While there is a larger disparity between the scores in the three Score Groups, overall, the scoring logic is able to classify essays at an acceptable level. The more promising outcome of this experiment was that the system was able to accurately identify all essays within the "Poor" Score Group, with a perfect Recall score of 1.

	Poor	Intermediate	Good	Average
Precision	0.94	0.64	0.60	0.73
Recall	1.00	0.50	0.69	0.73
F- Measure	0.97	0.56	0.64	0.72

Table 6.13: Average Scores for Precision, Recall and F-measure for Audience criterion

Based on the total number of essays used for testing and the total number of essays for which the human markers and the system returned the same Score Group, the exact agreement rate comes to 75%. Table 6.14 below shows the agreement rate for each of the scoring logics pertaining to the individual Score Groups while Table 6.13 shows the average Precision, Recall and F-measure scores across all Score Groups.

Score Group	System	Human Markers	Exact Agreement Rate
Poor	32	32	100%
Intermediate	14	28	50%
Good	22	30	73%
Overall	68	90	75%

Table 6.14: Exact Agreement rates for Audience criterion

Although there is still adjacent agreement when the system makes an error, wherein erroneously grouped essays are still within one Score Group rather than another (that is, no essays that are supposed to be in "Poor" are placed in "Good") the lower Recall value within the "Intermediate" Score Group is still a cause for concern.

Based on the value obtained via the Chi-square and t-test, the null hypothesis $\mathbf{H_0}$ is retained and the alternative $\mathbf{H_1}$ cannot be accepted. This allows us to conclude that there was insufficient evidence to show that there was a significant difference between the scores generated by the machine and those arrived at by human markers. It is thus concluded that the machine and human markers have a similar marking trend.

Following this, Chapters 7 to 9 will discuss the details of the remaining criteria, namely: Ideas, Character and Setting, and Cohesion. The format of these chapters will be largely similar to this one, thereby negating the need to repeat in full the methods used in the evaluations. The next chapter will describe the results and analysis of the Ideas criterion.

Chapter 7 - Group Scoring for Ideas

7.1 Introduction

As mentioned previously, the focus of this criterion is the creation and elaboration of ideas within the essay. As such, the focal point of this scoring logic will be based upon Events extracted from the Event Detection Stage. Events are taken to represent important happenings within a story; therefore, it is assumed that a well-written essay will contain a good number of Events with sufficient elaboration and a good Event Ratio, which is the number of Events over the total number of sentences.

Elaboration of Events in this instance refers to the amount of description within the essay. Here, the number of unique adjectives and adverbs are taken to represent, at surface level, the amount of elaboration that is present. Therefore, a high value for these features of text would give an essay the chance of being placed in a higher Score Group.

7.1.1 Overview of the Rubric Formalisation Process

The NAPLAN rubric consists of the following categories:

Band	Description
0	No evidence or insufficient evidence
1	Ideas are very few and very simple
2	Ideas are few but not elaborated

3	Ideas show some development or elaboration All ideas relate coherently to a central storyline
4	Ideas are substantial and elaborated Ideas effectively contribute to a central storyline The story contains a suggestion of an underlying theme
5	Ideas are generated, selected and crafted to explore a recognisable theme Ideas are skilfully used in the service of the storyline

Following the steps taken in Chapter 6, these categories were then sorted into 3 Score Groups as shown below:

	Poor	*Intermediate*	*Good*
Ideas	1-2	3	4-5

The features considered when placing an essay in its appropriate Score Group under Ideas are:

- Essay length
- Number of Events
- Event Ratio
- Number of unique adjectives
- Number of unique adverbs

Figures 7.1 and 7.2 below taken from Chapter 4 illustrate the Grouping and Score Grouping processes respectively.

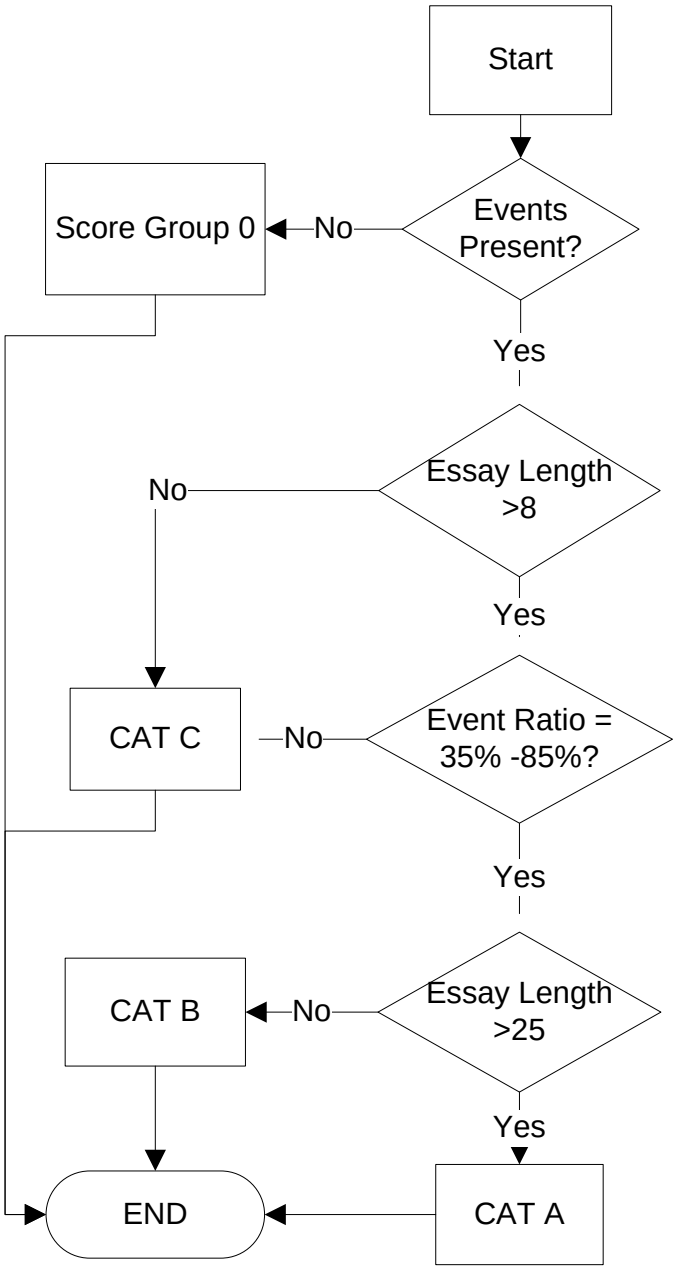

Figure 7.2: Grouping for Ideas, from Chapter 4

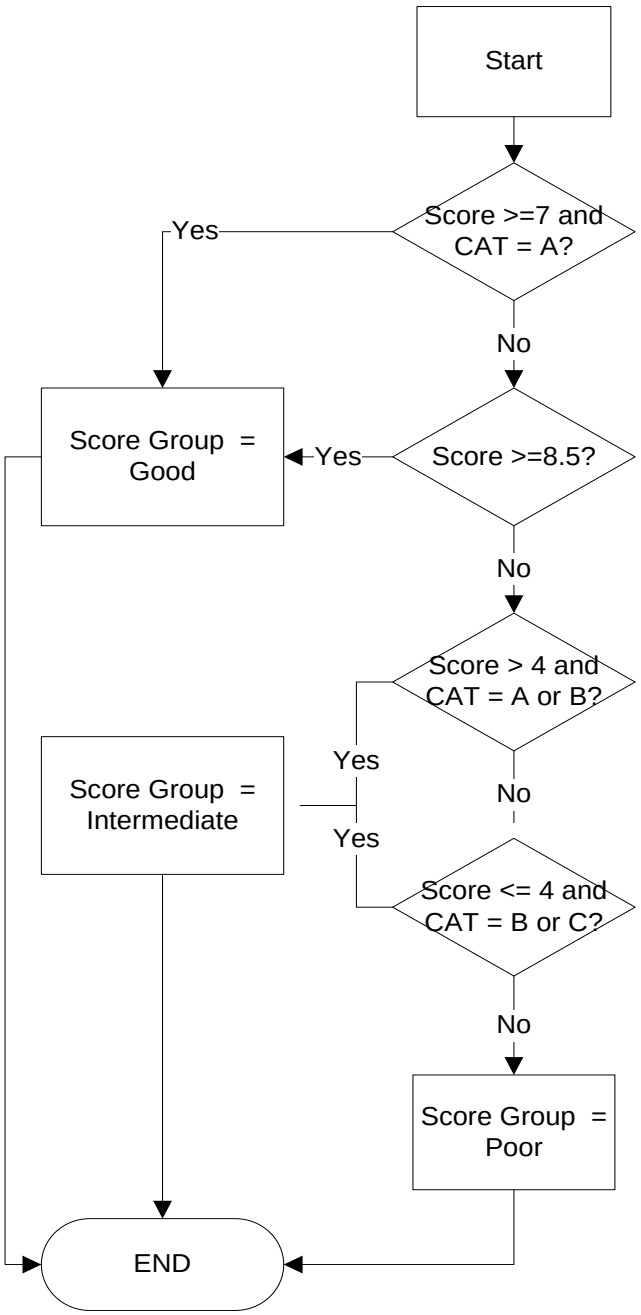

Figure 7.2: Score Grouping for Ideas, from Chapter 4

The main aim of this chapter is to ensure that the methods used for Score Grouping the essays are in fact valid. The rest of this chapter will first describe the methods used to carry out the experiments; this is followed by the hypothesis. The chapter then concludes with an analysis of the results gathered from those experiments.

7.2 Methodology Stage 1 - Precision, Recall, F-Measure and Exact Agreement Rate

Each essay from the test set was first sorted under its respective Score Group according to the band score it received from human markers. The 90 essays used for testing were then run through the scoring logic, and the results compared with those of the human markers. Table 7.1 shows each essay within the test set according to its respective band scores: Appendix F shows the full list of marks for all four criteria.

	Ideas Score Group		
Essay Name	*Poor*	*Intermediate*	*Good*
	ADKINS	ADANO	AMESS
	AGENBAG	AZMI	ANDREWS
	BEAVEN	BAKER.L	BAGIATIS
	BRIGGS	BENNETT	BAKER.C
	CHEREL	BERENTE	BELLIS
	CHETWYND	BERTOLA	BETTI
	COMBI	BIRSS	BOCCAMAZZO.C
	COYNE	BOCCAMAZZO.D	BOLES-RYAN
	DALE-FRASER	CATOVIC	BOTHMA
	DARCEY	COLBY	BOTH-WATSON
	DOPOE	CONN	BOWEN
	ELLERTON	DE PLEDGE	BRAMPTON
	FARLEY	DORRELL	BREAN
	FARRELL	GIANATTI	BYRNES
	FERGUSON	GORJY	CABUNALDA
	GUTHRIE	HAINES	CASTAING
	HAGUE	HIGGINSON	CHANDLER
	HALL	HOLT	CHARLES
	HANSEN.TA	JOHNSON	CHEDID
	HANSEN.TR	KARSKI	CHU
	HARLAND	KELLY	COPPARD
	HENRY	MASON	COWELL
	HODSON		CUNNINGHAM
	HUDSON		DANKS
	HUNTER		DIXON
	IOPPOLO		DORAN
	JONES		ESTENS
	MILTON		FOO
			FORWARD
			GALANTE
			GESTE
			HAWKETT
			HELSBY

			INGRAM KROLL LOH MAIN PALAYUKAN PERSSON VICKERY

Table 7.1: Individual Essay Score Groups according to Human Markers - Ideas

As with the previous experiment, this testing phase was conducted using a common test set of 90 essays chosen from a total of 189. Since the scoring logics were already trained using the other half of the data set comprised of 99 essays, there was no need to carry out the training phase again.

7.2.1 Experiments and Results

The objective of this experiment is largely the same as the one conducted for the previous criterion. As with the previous experiment, the objective here is to attempt to achieve a score as close to 1 as possible in the performance measures of Precision, Recall and F-Measure, while also trying to achieve as close to 100% in exact agreement.

The benchmark for the system to perform at an acceptable level is again set to a minimum average of 0.65 for the values of Precision, Recall and F-measure, while achieving no less than 0.55 in the overall exact agreement rate.

7.2.1.1 Score Group - Poor

Of the 90 essays used in this experiment, 28 essays were classified as belonging to the "Poor" Score Group based on their respective band scores. When put through the scoring logic, the system classified a total of 29 essays as belonging to this particular group, with 25 correctly classified essays and 9 errors.

Of these 9, 4 essays were incorrectly classified as “Poor” while 3 were incorrectly placed in the “Intermediate” Score Group as shown in Table 7.2.

Actual	System (True Positives)	Errors	
		False Positives	False Negatives
ADKINS	ADKINS	AZMI	AGENBAG
AGENBAG	BRIGGS	BENNETT	BEAVEN
BEAVEN	CHEREL	COLBY	DARCEY
BRIGGS	CHETWYND	KELLY	
CHEREL	COMBI		
CHETWYND	COYNE		
COMBI	DALE-FRASER		
COYNE	DOPOE		
DALE-FRASER	ELLERTON		
DARCEY	FARLEY		
DOPOE	FARRELL		
ELLERTON	FERGUSON		
FARLEY	GUTHRIE		
FARRELL	HAGUE		
FERGUSON	HALL		
GUTHRIE	HANSEN.TA		
HAGUE	HANSEN.TR		
HALL	HARLAND		
HANSEN.TA	HENRY		
HANSEN.TR	HODSON		
HARLAND	HUDSON		
HENRY	HUNTER		
HODSON	IOPPOLO		
HUDSON	JONES		
HUNTER	MILTON		
IOPPOLO			
JONES			
MILTON			

Table 7.2: Ideas Score Grouping results – Score Group “Poor”

7.2.1.1.1 Discussion

Using the results as input to the Precision, Recall and F-measure algorithms, we were able to arrive at the following results, shown in Table 7.3.

True Positives	False Positives	False Negatives	Precision	Recall	F - Measure
25	4	3	0.86	0.89	0.88

Table 7.3 Precision, Recall and F-Measure results for Ideas– Score Group "Poor"

The scoring logic for the "Poor" Score Group for the Ideas criterion was as expected, with high scores of over 0.85 across all performance measures. It was assumed that poorer essays would be much easier to identify since they would usually contain few to no Events, with extreme values in Event Ratios (either extremely low at 0-30% or extremely high with 85%-100%).

However, with the presence of 3 false negatives, where essays were incorrectly placed in a higher than "Poor" Score Group, it appears that the system does have some loopholes in the algorithm that might need to be addressed. In total, the system agreed with the human markers on 25 out of the 28 essays, giving this scoring logic an exact agreement rate of 89%, as described in Table 7.4.

Score Group	System	Human Makers	Agreement Rate
Poor	25	28	89.29%

Table 7.4: Agreement rate for Ideas – Score Group "Poor"

7.2.1.2 Score Group – Intermediate

For this Score Group, 22 essays were classified as Intermediate according to their band scores. According to the system however, 10 essays were correctly identified with a total of 27 errors. Of these 27 essays, 14 were incorrectly classified as "Intermediate"

while 13 essays which were supposed to be in this Score Group were classified otherwise.

Table 7.5 lists the essays correctly and incorrectly classified by the system when compared with the human makers, while Table 7.6 describes the values of the performance measures based on the results gathered.

Actual	System (True Positives)	Errors	
		False Positives	False Negatives
ADANO	BOCCAMAZZO.D	BAKER.C	ADANO
AZMI	BERTOLA	BOWEN	AZMI
BAKER.L	CATOVIC	BRAMPTON	BAKER.L
BENNETT	CONN	CHANDLER	BENNETT
BERENTE	ESTENS	CHARLES	BERENTE
BERTOLA	GORJY	DIXON	BIRRS
BIRSS	HAINES	FORWARD	COLBY
BOCCAMAZZO.D	HIGGINSON	GALANTE	DE PLEDGE
CATOVIC	JOHNSON	GESTE	DORRELL
COLBY	MASON	INGRAM	GIANATTI
CONN		PALAYUKAN	HOLT
DE PLEDGE		AGENBAG	KARSKI
DORRELL		BEAVEN	KELLY
GIANATTI		DARCEY	
GORJY			
HAINES			
HIGGINSON			
HOLT			
JOHNSON			
KARSKI			
KELLY			
MASON			

Table 7.5: Ideas Score Grouping results – Score Group "Intermediate"

7.2.1.2.1 Discussion

If the results shown in Table 7.6 are anything to go by, the scoring logic for this particular Score Group performs at an unacceptable level, thereby indicating the possibility that with regards to essays in the "Intermediate" Score Group, the features

mentioned are unrelated to an essay's grade. In an attempt to determine the reason for this unusually poor result, 2 essays were extracted for inspection namely CHU and INGRAM, excerpts from which are shown in Tables 7.7 and 7.8 respectively.

True Positives	False Positives	False Negatives	Precision	Recall	F – Measure
10	14	13	0.39	0.41	0.40

Table 7.6: Precision, Recall and F-measure results for Ideas - Score Group "Intermediate"

I wave goodbye to my best friend Nadine as I walk home. She has been my rock since my adoptive parents died. As I cried in her shoulder for a whole month since they died, we had become closer together than ever.

My journey home is not long. A brief 15 minute walk is all it takes. But these brief 15 minutes felt different. A little tickling on my neck gave me the feeling I was being watched. Something wasn't right. I hurried home as fast as I could. As soon as I got home, I locked all the doors and went upstairs to my room. Exhausted and scared, I collapsed onto the bed, fast asleep.

Table 7.7: Excerpt from sample essay - CHU

I was ridding my Bike with a Friend, we came to a stop when we Found a big woulden Box. WE took some time to have a Look the wondered what could be inside oF it, my Friend said "I dear you to see what's in there First he said" I Replied "NO WAY what if there is like u dead body or something?! Then I came up With a good Idea "what if we both have a look" Ok we both walked other there to see what could be inside the big would- an box we grabbed the lid and went to open it but there was a problem the lid was jamded so whe went back to my house and got some tools so we could open it. We Decided to walk there this time.

Table 7.8: Excerpt from sample essay - INGRAM

The former was placed in the "Good" Score Group as determined by its band score and the system also classified it as such. However, for the former, its band score also placed

it in the same group while the system classified it as "Intermediate" which adds to the false positive value of this particular scoring logic.

Upon closer inspection, it was found that the differences between the two essays were rather significant. The writing quality of the essay by CHU was substantially better than the other, but they both received the same band score of 4, which initially caused them to be placed in the same Score Group. However, when put through the scoring logic, the system classified CHU as the better essay, placing it in the "Good" Score Group while recognising that the essay by INGRAM was of a poorer quality in terms of the Ideas criterion, thus placing it in the "Intermediate" Score Group, which would have made more sense when comparing the two essays.

According to the results, the system agreed with the Score Groups for only 10 of the 22 essays scored by the human graders, giving an exact agreement rate of 45%, as shown in Table 7.9.

Score Group	System	Human Makers	Agreement Rate
Intermediate	10	22	45.45%

Table 7.9: Exact Agreement rate for Ideas – Score Group "Intermediate"

7.2.1.3 Score Group – Good

In the scoring logic used for this Score Group, the system fared much better than the former. According to human marker assigned scores, 40 essays belonged to this group. The scoring logic managed to correctly identify 28 of these essays while missing 11. An additional 9 essays were incorrectly classified as belonging to this Score Group. Table 7.10 shows the list of essays according to the results gathered from the test phase.

Actual		System (True Positives)	Errors	
			False Positives	False Negatives
AMESS	COWELL	AMESS	ADANO	BAKER.C
ANDREWS	CUNNINGHAM	ANDREWS	BAKER.L	BOWEN
BAGIATIS	DANKS	BAGIATIS	BERENTE	BRAMPTON
BAKER.C	DIXON	BAKER.C	BIRRS	CHANDLER
BELLIS	DORAN	BELLIS	DE PLEDGE	CHARLES
BETTI	ESTENS	BETTI	DORRELL	DIXON
BOCCAMAZZO.C	FOO	BOCCAMAZZO.C	GIANATTI	FORWARD
BOLES-RYAN	FORWARD	BOTHMA	HOLT	GALANTE
BOTHMA	GALANTE	BOTH-WATSON	KARSKI	GESTE
BOTH-WATSON	GESTE	BREAN		INGRAM
BOWEN	HAWKETT	BYRNES		PALAYUKAN
BRAMPTON	HELSBY	CABUNALDA		
BREAN	INGRAM	CASTAING		
BYRNES	KROLL	CHEDID		
CABUNALDA	LOH	CHU		
CASTAING	MAIN	COPPARD		
CHANDLER	PALAYUKAN	COWELL		
CHARLES	PERSSON	CUNNINGHAM		
CHEDID	VICKERY	DANKS		
CHU		DORAN		
COPPARD		FOO		
		HAWKETT		
		HELSBY		
		KROLL		
		LOH		
		MAIN		
		PERSSON		
		VICKERY		

Table 7.10: Ideas Score Grouping results – Score Group "Good"

7.2.1.3.1 Discussion

From the results gathered, the Precision, Recall and F-measure values were recorded as shown in Table 7.11.

True Positives	False Positives	False Negatives	Precision	Recall	F - Measure
28	9	11	0.76	0.72	0.74

Table 7.11: Precision, Recall and F-measure results for Ideas – Score Group "Good"

The performance of the scoring logic in this instance was at an acceptable level, with values of over 0.70 for all performance measures used. Although using the aforementioned features did allow us to identify most of the essays that belong in this particular Score Group, it is apparent that one cannot depend solely on features such as essay length and Event Ratios.

For example, essays that are of an unusually short length but are still of sound quality might get marked down a little more by the system than by its human counterparts, as in the case of the essay by BAKER.C.

Table 7.12 indicates that in terms of the exact agreement rate between the system and human markers, a total of 28 out of 40 essays were assigned to the same Score Group, giving a rate of 70%.

Score Group	System	Human Makers	Agreement Rate
Good	28	40	70.00%

Table 7.12: Exact Agreement rate for Ideas – Score Group "Good"

7.3 Methodology Stage 2 – Hypotheses Testing

Based on the features described in the conceptual framework for group scoring essays under the Ideas criteria, the following hypotheses were generated for this part of the thesis:

The null hypothesis $\mathbf{H_0}$ is:

> $\mathbf{H_0}$ **:** There is no significant difference between the human marker and machine-generated scores under the Ideas criterion.

The alternate hypothesis $\mathbf{H_1}$ would thus be:

$\mathbf{H_1}$ There is a significant difference between the human marker and machine-generated scores under the Ideas criterion.

7.3.1 Chi-Squared Goodness of Fit Test Results

Group	Observed	Expected
Poor	28	29
Intermediate	22	24
Good	40	37

Based on the data, the formula for determining the chi square was:

$$x^2 = \frac{(28-29)^2}{29} + \frac{(22-24)^2}{24} + \frac{(40-37)^2}{37}$$

$$x^2 = 0.03 + 0.016 + 0.24$$

$$x^2 = 0.44$$

DF	P=0.995	P=0.975	P=0.9	P=0.5	P=0.1	P=0.05	P=0.05	P=0.01	P=0.005
1	0.000	0.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838

For the null hypothesis to be rejected, the value of P should exceed 0.95. However, when compared with the highlighted row (DF 2) on chi-square distribution, the P value obtained is between 0.9 and 0.5, which is insufficient to reject the null hypothesis.

7.3.2 Paired T-Test Results

$$t = \frac{0.04-0}{\sqrt{0.57/(89)}}$$

$$t = 0.72$$

DF	0.1	0.05	0.02	0.01	0.005	0.002	0.001
89	1.6622	1.9870	2.3690	2.6322	2.8787	3.1844	3.4032

As shown above, the value of t should exceed the t critical value at 1.98 for there to be a significant difference between the machine and human marker scores. Therefore, since the t value obtained is much lower than the critical value at 0.72, the null hypothesis $\mathbf{H_0}$ cannot be rejected. This confirms that there is insufficient evidence to prove that the scores generated between the machine and human markers are significantly different.

7.4 Conclusion

Scoring an essay based entirely on the presence or absence of Events and the Event Ratio may be a viable way of scoring poorer and in most cases, good essays. It was assumed that separating the better essays from the rest of the group would be substantially harder since good essays often have rather subtle features that are hard to determine using conventional means. However, according to the results, the real problem lies in detecting essays that should belong to the “Intermediate” Score Group.

As discussed earlier, one of the main reasons for the poor performance with regards to that scoring logic is the large inconsistency within the human marker assigned scores. Often, an essay of an apparently poorer quality would be marked the same as an essay actually deserving of that score, which would affect the calibration of the system as a whole.

	Poor	Intermediate	Good	Average
Precision	0.86	0.39	0.76	0.67
Recall	0.89	0.41	0.72	0.67
F- Measure	0.88	0.40	0.74	0.67

Table 7.13: Average Scores for Precision, Recall and F-measure for Ideas criterion

Score Group	System	Human Makers	Agreement Rate
Poor	25	28	89.29%
Intermediate	10	22	45.45%
Good	28	40	70.00%
Overall	**63**	**90**	**70.00%**

Table 7.14: Agreement rates for Ideas criterion

With the total number of errors coming to 27 out of 90 essays, the system achieved an agreement rate of 63% with the human markers, as shown in Table 7.14. Overall, the system managed to achieve an average of 0.67 across all performance measures as shown in Table 7.13, which allows us to conclude that the presence of Events within an essay, together with the Event Ratio, does have a bearing on the score an essay receives.

Although 27 essays were incorrectly classified, the system still managed to achieve an adjacent agreement with its human counterparts, with no essays placed more than one Score Group apart.

From the results, it appears that having an Event interpreted as the presence of an idea within an essay seems like a promising method of essay grading, with an overall performance of over 0.65 for Precision, Recall and F-measure. This allows us to conclude that Events, together with other features of a narrative essay such as unique adjectives and adverbs, are related to the score it receives.

However, given that the performance of the "Intermediate" scoring logic is rather low, it might appear that the system is more lenient in some aspects when compared with a human marker. Of the 25 essays placed in this group by the system, 11 essays were scored lower while 3 were scored higher. However, having stated this, the performance of the other two scoring logics make up for this by achieving an F-measure of 0.88 and 0.74 in the "Poor" and "Good" Score Groups respectively.

Based on the results gathered from the Chi-Square and T tests, the null hypothesis $\mathbf{H_0}$ is kept and the alternative $\mathbf{H_1}$ is rejected.

The next chapter, Chapter 7, focuses on the next criterion, Character and Setting.

Chapter 8 - Group Scoring for Character and Setting

8.1 Introduction

The basic concept of the scoring process for this criterion is assumed to be linked somewhat to the Ideas criterion, since scoring it also relies on the presence of Events within the essay, though not entirely dependent on it.

8.1.1 Overview of the Rubric Formalisation Process

For the Character and Setting criterion, the NAPLAN Rubric has the following categories:

Band	Description
0	No evidence or insufficient evidence
1	Only names the characters or gives their roles (e.g. father, the teacher, my friend, dinosaur, we, Jim) and/or Only names the setting (e.g. school, the place we were at); setting is vague or confused
2	Suggestion of characterisation through brief descriptions or speech or feelings, but lacks substance or continuity and/or Suggestion of setting through very brief or superficial descriptions of place and/or time
3	Characterisation emerges through descriptions, action, speech or the

	attribution of thoughts and feelings to a character and/or Setting emerges through the description of place, time and atmosphere
4	Effective characterisation. Details are selected to create distinct characters and/or Maintains a sense of setting throughout. Details are selected to create a sense of place and atmosphere

A mentioned earlier in Chapter 4, the difference between the Ideas and Character and Setting criteria is that the development of the characters within and/or the setting (State) depicted has more bearing on the essay's score, rather than the Events that encompass them.

Hence, the scoring conditions for the Ideas criterion are taken and applied to this one albeit with some modifications to the focus. The Character and Settings criterion was earlier separated into the three groups as follows:

	Poor	*Intermediate*	*Good*
Character and Setting	1-2	3	4

The features considered when placing an essay in its appropriate Score Group under Ideas are:

- Essay length
- Number of Events

- Event Ratio
- Number of unique adjectives
- Number of unique adverbs
- Physical and/or Mental State

Taken from Chapter 4, Figures 8.1and 8.2 illustrate the Grouping and Score Grouping processes respectively.

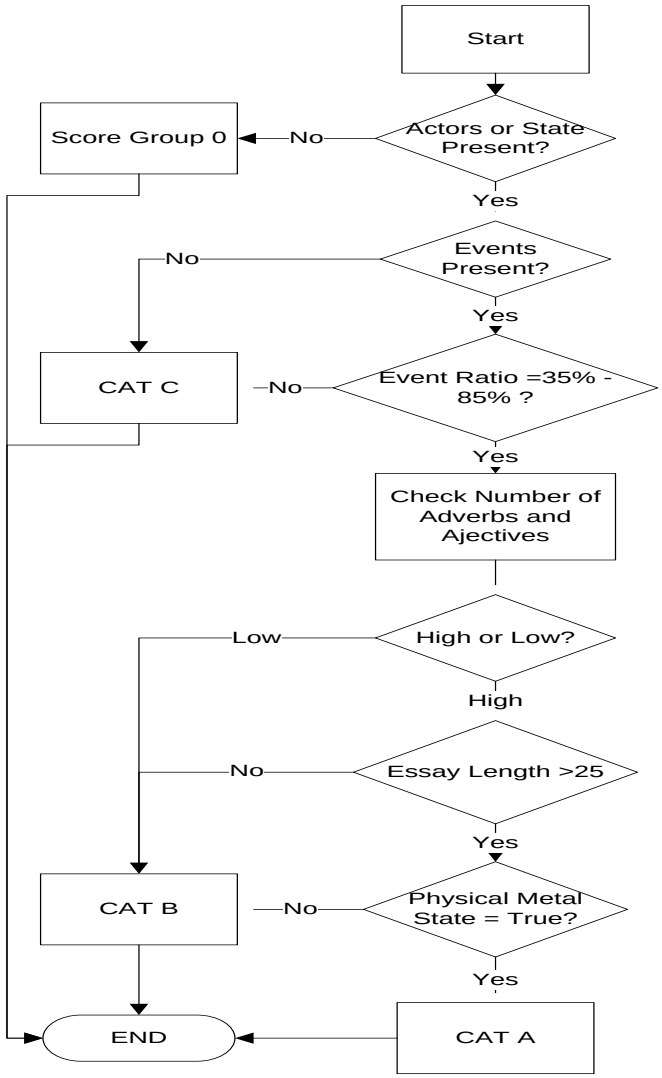

Figure 8.1: Grouping for Character and Setting, from Chapter 4

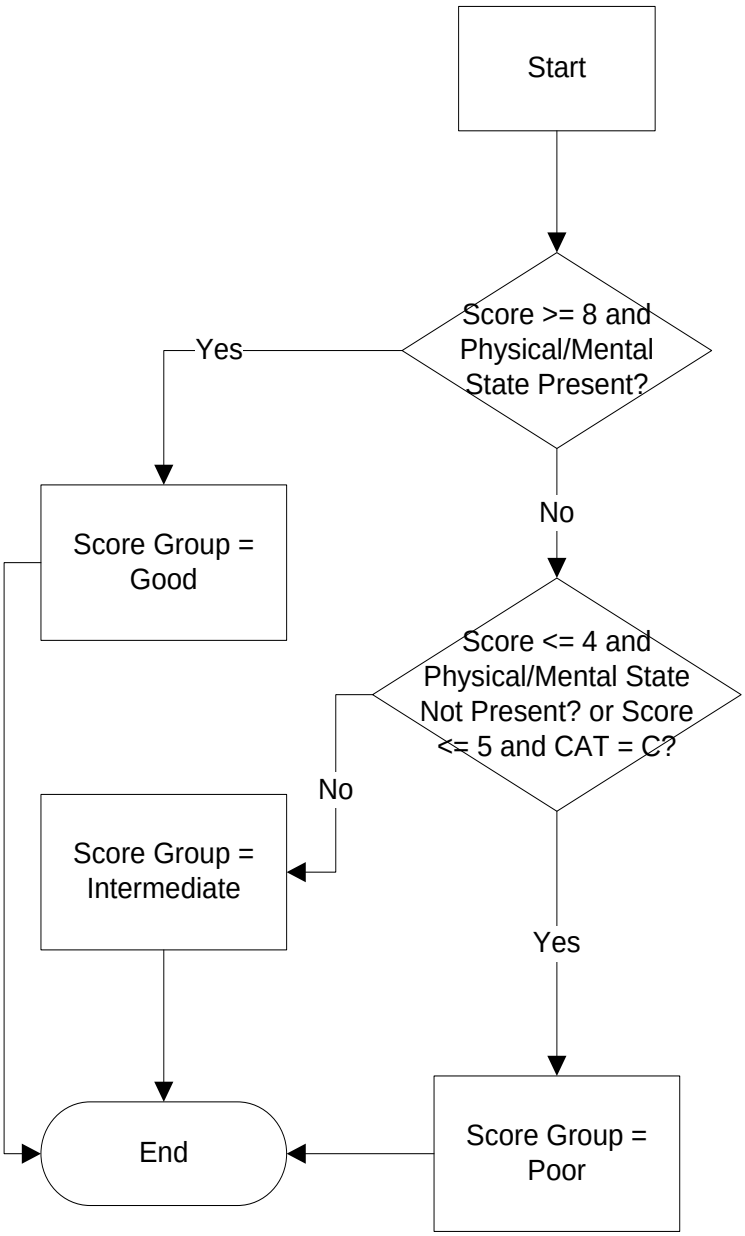

Figure 8.2: Score Grouping for Character and Setting, from Chapter 4

This chapter aims to ensure that the method used in Score Grouping the essays is in fact a valid one. Towards this end, the rest of this chapter will detail the steps used to carry out the experiments, followed by the hypothesis which tests the validity of the methods. The chapter then concludes with an analysis of the experiments' results.

8.2 Methodology Stage 1 – Precision, Recall, F-Measure and Exact Agreement Rate

By now, the method used to carry out the tests should be quite familiar. The 90 essays within the test set were first sorted into their Score Groups based on their respective band scores as shown in Table 8.1. Something of note that bears repeating is that some essays might score particularly well in one criterion while surprisingly poorly in another.

An example of this is the essay by CHU, which performed fairly well for the Ideas criterion according to marks allocated by human markers, but in contrast, fared rather poorly for the Character and Setting criterion. In all, there were 41 essays that were sorted into the "Poor" Score Group, with 30 and 19 essays sorted into the "Intermediate" and "Good" Score Groups respectively. For the full list of band scores assigned to all essays, please refer to Appendix F.

	Character and Setting Score Group		
Essay Name	*Poor*	*Intermediate*	*Good*
	ADKINS	ADANO	BAKER.C
	AGENBAG	AMESS	BELLIS
	AZMI	ANDREWS	BOWEN
	BEAVEN	BAGIATIS	BYRNES
	BENNETT	BAKER.L	CATOVIC
	BERTOLA	BERENTE	CHANDLER
	BIRSS	BETTI	CHEDID
	BRIGGS	BOCCAMAZZO.D	COPPARD
	CHEREL	BOCCAMAZZO.C	COWELL
	CHETWYND	BOLES-RYAN	DANKS
	CHU	BOTHMA	DORAN
	COMBI	BOTH-WATSON	FOO
	CONN	BRAMPTON	FORWARD
	COYNE	BREAN	GESTE
	DALE-FRASER	CABUNALDA	KROLL
	DARCEY	CASTAING	LOH
	DE PLEDGE	CHARLES	MAIN
	DOPOE	COLBY	PALAYUKAN
	DORRELL	CUNNINGHAM	PERSSON

	ELLERTON FARLEY FARRELL FERGUSON GORJY GUTHRIE HAGUE HALL HANSEN.TA HANSEN.TR HARLAND HENRY HIGGINSON HODSON HUDSON HUNTER IOPPOLO JONES KELLY MASON MILTON VICKERY	DIXON ESTENS GALANTE GIANATTI HAINES HAWKETT HELSBY HOLT INGRAM JOHNSON KARSKI	

Table 8.1: Individual Essay Score Groups according to Human Markers – Character and Setting

8.2.1 Experiments and Results

With the average performance benchmark across all performance measures set at a minimum of 0.65, evaluations were carried out using the Precision, Recall and F-measure values. As before, a 100% exact agreement rate is worked towards while ensuring the system fares no less than 55%. The following sections will describe in further detail the outcomes of the tests.

8.2.1.1 Score Group - Poor

Using scores assigned by a human marker to initially group the essays, 41 essays were allocated to this Score Group. As shown in Table 8.2, of these 41 essays, 35 were identified by the system as belonging to the same group, with 7 essays erroneously placed in other Score Groups.

Actual	System (True Positives)	Errors	
		False Positives	False Negatives
ADKINS	ADKINS	BAKER.C	BERTOLA
AGENBAG	AGENBAG	CHANDLER	BIRSS
AZMI	AZMI	COLBY	CHU
BEAVEN	BEAVEN	INGRAM	CONN
BENNETT	BENNETT		DE PLEDGE
BERTOLA	BRIGGS		DORRELL
BIRSS	CHEREL		HIGGINSON
BRIGGS	CHETWYND		VICKERY
CHEREL	COMBI		
CHETWYND	COYNE		
CHU	DALE-FRASER		
COMBI	DARCEY		
CONN	DOPOE		
COYNE	ELLERTON		
DALE-FRASER	FARLEY		
DARCEY	FARRELL		
DE PLEDGE	FERGUSON		
DOPOE	GORJY		
DORRELL	GUTHRIE		
ELLERTON	HAGUE		
FARLEY	HALL		
FARRELL	HANSEN		
FERGUSON	HANSEN		
GORJY	HARLAND		
GUTHRIE	HENRY		
HAGUE	HODSON		
HALL	HUDSON		
HANSEN.TA	HUNTER		
HANSEN.TR	IOPPOLO		
HARLAND	JONES		
HENRY	KELLY		
HIGGINSON	MASON		
HODSON	MILTON		
HUDSON			
HUNTER			
IOPPOLO			
JONES			
KELLY			
MASON			
MILTON			
VICKERY			

Table 8.2: Character and Setting Score Grouping results – Score Group "Poor"

8.2.1.1.1 Discussion

As with the test conducted previously, the scoring logic for the "Poor" Score Group shares a high exact agreement rate with the human markers, as indicated in Table 8.3, with 35 essays classified similarly to the human marker assigned scores. This result shows that as far as poor essays go, the characteristics that would place them in the "Poor" Score Group for one particular criterion would probably result in those essays being placed in the same Score Group for other criteria.

Score Group	System	Human Makers	Agreement Rate
Poor	35	41	85.37%

Table 8.3: Exact Agreement rate for Character and Setting – Score Group "Poor"

The relatively low number of false negatives, although usually a good sign, does signify that the system is stricter in the scoring than are its human counterparts, as is the case with the other scoring logics pertaining to the other criteria.

True Positives	False Positives	False Negatives	Precision	Recall	F - Measure
35	4	8	0.90	0.81	0.85

Table 8.4: Precision, Recall and F-measure results for Character and Setting – Score Group "Poor"

The results indicate that this section of the scoring logic shows promise, with high values in Precision, Recall and F-measure at 0.90, 0.81 and 0.85 respectively, as shown in Table 8.4.

8.2.1.2 Score Group – Intermediate

Of the 30 essays that belong to this Score Group when sorted according to the band scores assigned by the human marker, the scoring logic identified 25 essays that belong to the same group. A total of 20 essays were placed in this group by the scoring logic, while 4 essays that were supposed to be in this group according to the human marker scores were placed in the "Good" Score Group.

Of the 15 false positives, 7 essays were supposed to be allocated to the "Poor" Score Group, while 8 belonged to the "Good" Score Group. Table 8.5 shows the list.

Actual	System (True Positives)	Errors	
		False Positives	False Negatives
ADANO AMESS ANDREWS BAGIATIS BAKER.L BERENTE BETTI BOCCAMAZZO.D BOCCAMAZZO.C BOLES-RYAN BOTHMA BOTH-WATSON BRAMPTON BREAN CABUNALDA CASTAING CHARLES COLBY CUNNINGHAM DIXON ESTENS GALANTE GIANATTI HAINES HAWKETT HELSBY HOLT INGRAM JOHNSON KARSKI	ADANO AMESS ANDREW BERENTE BETTI BOCCAMAZZO.D BOCCAMAZZO.C BOTH-WATSON BRAMPTON BREAN CABUNALDA CHARLES CUNNINGHAM DIXON ESTENS GALANTE GIANATTI HAWKETT JOHNSON KARSKI	BERTOLA BOWEN BYRNES CATOVIC CHU CONN DANKS DE PLEDGE DORAN DORRELL FORWARD GESTE HIGGINSON MAIN VICKERY	BAGIATIS BAKER.L BOLES-RYAN BOTHMA CASTAING COLBY HAINES HELSBY HOLT INGRAM

Table 8.5: Character and Setting Score Group results – Score Group "Intermediate"

8.2.1.2.1 Discussion

Once again, the scoring logic was shown to be quite a bit stricter in the grading process than were its human counterparts, with disagreements on 20 essays, 13 of which were placed in a poorer Score Group by the system. One of the reasons for this might be that human markers are able to perform a much deeper, though sometimes biased, contextual analysis of the text.

Obviously, one of the objectives of the scoring logic is to provide some contextual analysis using the surface features of the text but the difference shows when the results are observed; this difference in ability was thought to be the reason for the disparity.

However, another point worthy of consideration is one that has come to light before, which is the inconsistency in scores assigned by human markers. When two essays are placed within the same Score Group according to the band score they receive from the human markers, one would assume that those two essays would be of a similar quality. If not, then at least there should be little that sets them apart, or that an essay of the highest quality in the "Poor" Score Group should not be better than the essay of the lowest quality within the "Intermediate" Score Group.

Thus, if this were not the case, then there would be little to no consistency within the marks assigned, which then begs the question of whether those scores are a correct reflection of the essay's quality in the first place. One such example of inconsistency is

shown using again the essay by CHU, but this time in comparison with an essay by CASTAING.

I wave goodbye to my best friend Nadine as I walk home. She has been my rock since my adoptive parents died. As I cried in her shoulder for a whole month since they died, we had become closer together than ever. My journey home is not long. A brief 15 minute walk is all it takes. But these brief 15 minutes felt different. A little tickling on my neck gave me the feeling I was being watched. Something wasn't right. I hurried home as fast as I could. As soon as I got home, I locked all the doors and went upstairs to my room. Exhausted and scared, I collapsed onto the bed, fast asleep.

Table 8.6: Excerpt from sample essay - CHU

I looked at and waited anxiously, and also scared. Finally, Mum and Dada burst through the door. "Aw our darling daughter, how are you feeling today?" Mum asked. "Yeah fine", I replied. The actual truth was that I felt the same as every day. Maybe something was different about today though? What was I thinking? That's what I thought every day and now was no different. I stared down at my frail body as I sat in the hospital bed. I longed for my old body, my old long hair, my old home but most of all my old life. Ever since I was diagnosed with Leukaemia everything and everyone had changed. Life was tough but it was also weird. I just wished I could be perfect. I lay my head on the pillow and tried hard not to cry.

Table 8.7: Excerpt from sample essay - CASTAING

Upon reading the two excerpts from these essays, most might agree that the quality of the former is not far off from the latter; however, the former was classified as belonging to the "Poor" Score Group according to the band score it received from human markers, whereas the latter was one group higher. Perhaps these discrepancies indicated by the large number of false positives, are due to the inconsistency among human markers, rather than the result of system logic error.

True Positives	False Positives	False Negatives	Precision	Recall	F - Measure
20	15	10	0.57	0.67	0.62

Table 8.8: Precision, Recall and F-measure results for Character and Setting – Score Group "Intermediate"

Having said that, the system still achieves a score of 0.57 in Precision, while faring slightly better in Recall and F-measure values with scores of 0.67 and 0.62 respectively, as detailed in Table 8.8. This result far surpasses expectations since results for this Score Group from previous tests have showed much poorer results. The number of essays that were in both scoring processes placed in the same Score Group was 20 out of the 30 essays grouped by the human markers, thus giving a more promising exact agreement rate of 66.67%, as shown in Table 8.9.

Score Group	System	Human Makers	Exact Agreement Rate
Intermediate	20	30	66.67%

Table 8.9: Exact Agreement rate for Character and Setting – Score Group "Intermediate"

8.2.1.3 Score Group – Good

As shown in Table 8.10, the scoring logic for this Score Group shares little agreement with the human marker scores. Of the 20 essays placed in this group according to their respective band scores, the system identified only 6, with a total of 15 disagreements. Of these 15 essays, 4 were placed in a higher Score Group while 11 were placed in a lower one.

Actual	System (True Positives)	Errors	
		False Positives	False Negatives
BAKER.C BELLIS BOWEN BYRNES CATOVIC CHANDLER CHEDID COPPARD COWELL DANKS DORAN FOO FORWARD GESTE KROLL LOH MAIN PALAYUKAN PERSSON	BELLIS CHEDID COPPARD COWELL KROLL LOH PALAYUKAN PERSSON	BAGIATIS BAKER.L BOLES-RYAN BOTHMA CASTAING HAINES HELSBY HOLT	BAKER.C BOWEN BYRNES CATOVIC CHANDLER DANKS DORAN FORWARD GESTE MAIN

Table 8.10: Character and Setting Score Grouping results – Score Group "Good"

8.2.1.3.1 Discussion

Taking into account the performance of the previous scoring logics with regards to this Score Group, the performance of the scoring logic here was unexpectedly poor. As can be seen in Table 8.11, the highest score achieved across all performance measures was

0.50 in Precision, which is sub-par in itself. The system fared even worse in Recall and F-measure scores, obtaining values of only 0.47 and 0.49 respectively.

True Positives	False Positives	False Negatives	Precision	Recall	F - Measure
9	9	10	0.50	0.47	0.49

Table 8.11: Precision, Recall and F-measure results for Character and Setting – Score Group "Good"

With 9 false negatives, it could be said that the system is substantially more stringent in the scoring process than are the human markers. One reason for this rather large disparity could be that when marking an essay, human markers are specifically told not to penalise an essay for the same recurring mistake. For example, if an essay is marked down for not having sufficient detail or elaboration within an Event, it cannot be penalised again for the same thing when it is considered in another criterion even though they may share similar concepts. Such is the case between the Ideas and Character and Setting criteria.

Therefore, it is also assumed by the system that the opposite is true, where an essay would not receive credit for the same feature twice. This is addressed by the weighted scoring system discussed in the conceptual framework, where the same features would receive a different weighting for different criteria. It would appear however, that this method of scoring does not tie in well when it comes to the agreement between the system and human markers, as can be seen when considering the results in Table 8.12.

Score Group	System	Human Makers	Exact Agreement Rate
Good	9	19	47.37%

Table 8.12: Exact Agreement rate for Character and Setting – Score Group "Good"

8.3 Methodology Stage 2 -Hypothesis Testing

Based on the features described in the conceptual framework for group scoring essays under the Character and Settings criterion, the following hypotheses were generated for this part of the thesis:

The null hypothesis, $\mathbf{H_0}$ is:

> $\mathbf{H_0}$ **:** There is no significant difference between the human marker scores and the machine-generated scores for the Character and Settings criterion.

The alternate hypothesis, $\mathbf{H_1}$ would thus be:

> $\mathbf{H_1}$: There is a significant difference between the human marker scores and the machine-generated scores for the Character and Settings criterion.

8.3.1 Chi-Squared Goodness of Fit Test Results

Group	Observed	Expected
Poor	39	41
Intermediate	35	30
Good	18	19

Based on the data, the formula for determining the Chi square was:

$$x^2 = \frac{(39-41)^2}{41} + \frac{(35-30)^2}{30} + \frac{(18-19)^2}{19}$$

$$x^2 = 00.97 + 0.833 + 0.052$$

$$x^2 = 0.98$$

DF	P=0.995	P=0.975	P=0.9	P=0.5	P=0.1	P=0.05	P=0.05	P=0.01	P=0.005
1	0.000	0.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838

As with the previous experiments, for the null hypothesis to be rejected, the value of P should exceed 0.95. Referring to the Table above, the P value obtained is once again between 0.9 and 0.5, which means that the null hypothesis cannot be rejected.

8.3.2 Paired T-Test Results

$$t = \frac{0.03-0}{\sqrt{0.64/(89)}}$$

$$t = 0.48$$

DF	0.1	0.05	0.02	0.01	0.005	0.002	0.001
89	1.6622	1.9870	2.3690	2.6322	2.8787	3.1844	3.4032

For there to be a significant difference between the machine and human marker scores, the value of t should exceed the t critical value which is at 1.98. As shown in the Table above, the t value obtained is lower than the critical value at 0.48.

Thus, the null hypothesis $\mathbf{H_0}$ cannot be rejected which shows that there is insufficient evidence to indicate a significant difference between the scores generated by the machine and those by human markers.

8.4 Conclusion

This chapter detailed the performances of each of the scoring logics for the Character and Setting criterion. The scoring logic for this criterion differs slightly from the one used for the Ideas criterion since the presence of an Event, while playing a large part, does not entirely define the Score Group to which an essay should belong. Instead, the focus is on the characters themselves and the setting of the story, as established by the author.

Using these conditions as a guide, it is assumed that with a high number of unique adjectives and adverbs, together with the presence of a Physical or Mental State pertaining to the characters, it would be possible to correctly place an essay in its relevant Score Group. Judging from the results gathered, it appears that while the scoring logic performs at an acceptable level in accordance with the agreement rate between with the human markers, the system seems to produce a large disparity when it comes to placing essays in the “Good” Score Group.

This result stands out from the previous tests in that it is for this particular Score Group that the system performs poorly, where normally we would have seen a greater disagreement between the system and the human markers for the “Intermediate” Score Group.

	Poor	Intermediate	Good	Average
Precision	0.90	0.57	0.50	0.66
Recall	0.81	0.67	0.47	0.65
F- Measure	0.85	0.62	0.49	0.65

Table 8.13: Average Scores for Precision, Recall and F-measure for Character and Setting criterion

However, considering the overall performance of the scoring logic for this criterion, the system performed relatively well, achieving average scores of 0.65 across the metrics of Precision, Recall and F-measure although individually the scoring logic for the "Good" Score Group fares rather poorly.

Score Group	System	Human Makers	Exact Agreement Rate
Poor	35	41	85.37%
Intermediate	20	30	66.67%
Good	9	19	47.37%
Overall	64	90	71.11%

Table 8.14: Exact Agreement rates for Character and Setting criterion

From the results shown in Tables 8.13 and 8.14, it can be affirmed that the attributes and features identified earlier in this chapter are somewhat relevant to the score an essay receives. In addition, from the results gathered from the Chi square and T tests, the null hypothesis, $\mathbf{H_0}$ cannot be rejected and thus the alternative,$\mathbf{H_1}$ is rejected.

Next Chapter 9 will discuss the final criterion considered for this thesis, which is Cohesion.

Chapter 9 - Group Scoring for Cohesion

9.1 Introduction

As mentioned in Chapter 4 and according to the marking rubric, the main focus of this criterion is the "control of multiple thread and relations" within the essay. The rubric measures this ability through the use of word association, substitution and other referring words.

9.1.1Overview of the Rubric Formalisation Process

The NAPLAN rubric describes the criterion for Cohesion as:

Band	Description
0	Symbols or drawings
1	Links are missing or incorrect Short script Often confusing for the reader
2	Some correct links between sentences (do not penalise for poor punctuation) Most referring words are accurate
3	Cohesive devices are used correctly to support reader understanding Accurate use of referring words Meaning is clear and text flows well in a sustained piece of writing

4	A range of cohesive devices is used correctly and deliberately to enhance reading An extended, highly cohesive piece of writing showing continuity of ideas and tightly linked sections of text

In this thesis, it is assumed that the number of unique connectives, simple and advanced, would relate to the above-mentioned features, together with the Events and other various characteristics. Therefore, the following were considered when placing an essay in its appropriate Score Group for the Cohesion criterion:

- Essay length
- Number of Events
- Event Ratio
- Number of simple connectives
- Number of advanced connectives

The Cohesion criterion was earlier separated into the following 3 groups:

	Poor	*Intermediate*	*Good*
Cohesion	1-2	3	4

Figures 9.1 and 9.2 further illustrate the Grouping and Score Grouping logics respectively.

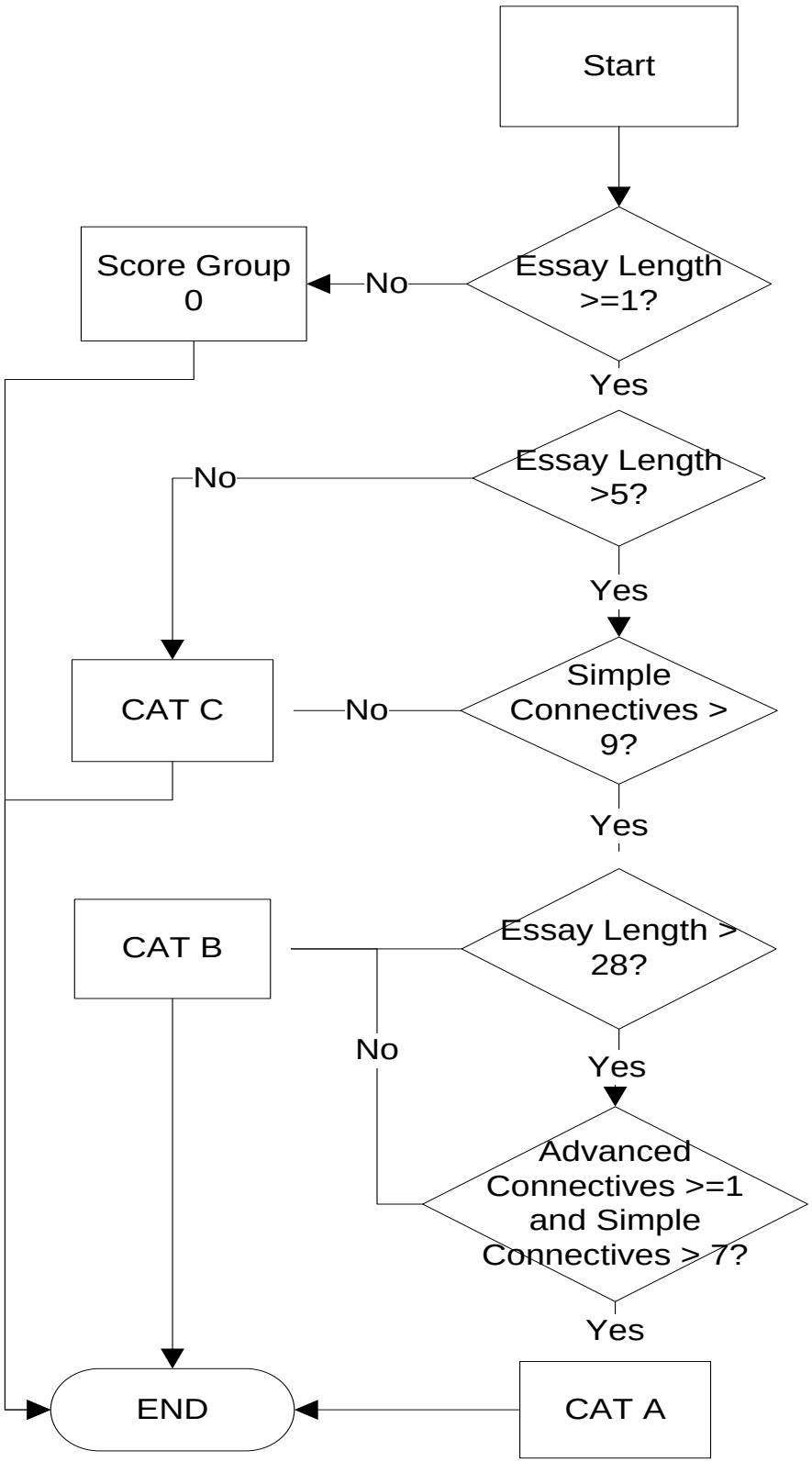

Figure 9.1: Grouping for Cohesion, from Chapter 4

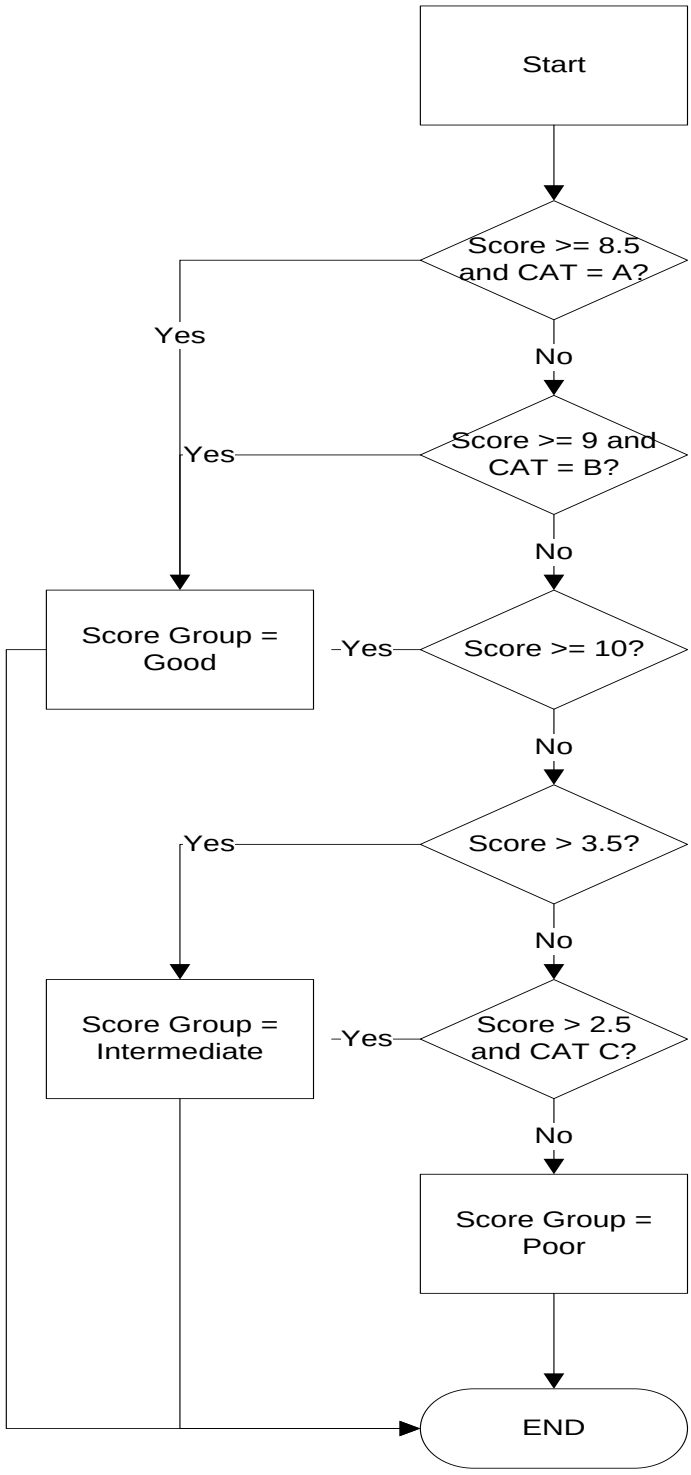

Figure 9.2: Score grouping for Cohesion, from Chapter 4

This chapter aims to provide validation that the method used in Score Grouping the essays under the Cohesion criterion is in fact a valid one. The next sections will firstly

state the hypothesis which tests the validity of the grouping process, followed by an explanation of the experiments carried out and lastly an analysis of the results gathered from those experiments.

9.2 Methodology Stage 1 - Precision, Recall, F-Measure and Exact Agreement Rate

Adhering to the methods used in the previous three tests, the same 90 essays were run through the scoring logic after being sorted into their respective Score Groups in accordance with the marks the received from human markers for this criterion.

The resultant groupings are shown in Table 9.1, with a total of 34 and 19 essays being placed in the "Poor" and "Good" Score Groups respectively. The number of essays placed in "Intermediate" Score Group was unusually high, with 37 essays.

This was in contrast to the previous tests where it was usually one of the other two Score Groups which contained the bulk of essays. Once again, for the full list of band scores assigned to all essays, please refer to Appendix F.

	Cohesion Score Group		
Essay Name	*Poor*	*Intermediate*	*Good*
	ADKINS	ADANO	BELLIS
	AGENBAG	AMESS	BOTH-WATSON
	AZMI	ANDREWS	BRAMPTON
	BEAVEN	BAGIATIS	BREAN
	BENNETT	BAKER.L	BYRNES
	BERTOLA	BAKER.C	CASTAING
	BRIGGS	BERENTE	CATOVIC
	CHEREL	BETTI	CHANDLER
	CHETWYND	BIRSS	CHEDID
	COMBI	BOCCAMAZZO.D	COPPARD
	CONN	BOCCAMAZZO.C	DANKS

	COYNE	BOLES-RYAN	DORAN
	DALE-FRASER	BOTHMA	FOO
	DARCEY	BOWEN	KARSKI
	DOPOE	CABUNALDA	KROLL
	ELLERTON	CHARLES	LOH
	ESTENS	CHU	MAIN
	FARLEY	COLBY	PERSSON
	FARRELL	COWELL	VICKERY
	FERGUSON	CUNNINGHAM	
	GUTHRIE	DE PLEDGE	
	HAGUE	DIXON	
	HALL	DORRELL	
	HANSEN.TA	FORWARD	
	HANSEN.TR	GALANTE	
	HARLAND	GESTE	
	HENRY	GIANATTI	
	HODSON	GORJY	
	HUDSON	HAINES	
	HUNTER	HAWKETT	
	IOPPOLO	HELSBY	
	JONES	HIGGINSON	
	KELLY	HOLT	
	MILTON	INGRAM	
		JOHNSON	
		MASON	
		PALAYUKAN	

Table 9.1: Individual Essay Score Groups according to Human Markers - Cohesion

9.2.1 Experiments and Results

Following the benchmarks of the previous tests, the minimum acceptable values for Precision, Recall and F-measure were set to 0.65, while attempting to achieve as close to 0.91 as possible. The exact agreement rate is once again expected to be no lower than 55%.

9.2.1.1 Score Group - Poor

Of the 34 essays sorted into this Score Group based on scores assigned by their human marker, the system agreed on 30 essays, with no false positives and only 4 false negatives. The 2 essays that were placed in this group by human marker scores were

instead sorted into the higher, “Intermediate” Score Group. Table 9.2 shows the details.

Actual	System (True Positives)	Errors	
		False Positives	False Negatives
ADKINS	ADKINS	NIL	BENNETT
AGENBAG	AGENBAG		BERTOLA
AZMI	AZMI		CONN
BEAVEN	BEAVEN		ESTENS
BENNETT	BRIGGS		
BERTOLA	CHEREL		
BRIGGS	CHETWYND		
CHEREL	COMBI		
CHETWYND	COYNE		
COMBI	DALE-FRASER		
CONN	DARCEY		
COYNE	DOPOE		
DALE-FRASER	ELLERTON		
DARCEY	FARLEY		
DOPOE	FARREK		
ELLERTON	FERGUSON		
ESTENS	GUTHRIE		
FARLEY	HAGUE		
FARRELL	HALL		
FERGUSON	HANSEN.TA		
GUTHRIE	HANSEN.TR		
HAGUE	HARLAND		
HALL	HENRY		
HANSEN.TA	HODSON		
HANSEN.TR	HUDSON		
HARLAND	HUNTER		
HENRY	IOPPOLO		
HODSON	JONES		
HUDSON	KELLY		
HUNTER	MILTON		
IOPPOLO			
JONES			
KELLY			
MILTON			

Table 9.2: Cohesion Score Grouping results – Score Group “Poor”

9.2.1.1.1 Discussion

As expected from results based on the previous tests, the section of the scoring logic pertaining to this Score Group performed very well, with a perfect score of 1 in terms of Precision, indicating that no essays were placed in this Score Group where they should have been otherwise placed.

Test scores for Recall and F-measure also returned high values, with scores of 0.88 and 0.94 respectively. Judging from the results, we can ascertain that while the relationship between the presences of the aforementioned textual features, together with the Events within the essay has not yet been established, the absence of these features is definitely significant in determining whether or not an essay should be placed in a higher Score Group.

True Positives	False Positives	False Negatives	Precision	Recall	F - Measure
31	0	2	1	0.88	0.94

Table 9.3: Precision, Recall and F-Measure results for Cohesion – Score Group "Poor"

With 30 out the 34 essays agreed upon between the system and the human markers, the exact agreement rate came to 88% as Table 9.4 shows.

Score Group	System	Human Makers	Exact Agreement Rate
Poor	30	34	88.24%

Table 9.4: Exact Agreement Rate for Audience – Score Group Cohesion

9.2.1.2 Score Group – Intermediate

This Score Group, according to the human marker scores contained the majority of the test set, with 37 essays in this group. Of these 37 essays, the system allocated 26 to

the same Score Group, with 12 false negatives, which were all found to be placed in the "Good" Score Group by the system.

Of the 11 false positives, 7 essays belonged to the "Good" Score Group while 4 were placed in a group lower than that determined by the human marker scores. Table 9.5 shows the details.

Actual	**System (True Positives)**	**Errors**	
		False Positives	**False Negatives**

ADANO	ADANO	BENNETT	AMESS
AMESS	ANDREWS	BERTOLA	BAKER.L
ANDREWS	BAGIATIS	BOTH-WATSON	BIRSS
BAGIATIS	BAKER.C	CASTAING	BOCCAMAZZO.D
BAKER.L	BERENTE	CATOVIC	BOLES-RYAN
BAKER.C	BETTI	CHANDLER	BOTHMA
BERENTE	BOCCAMAZZO.C	CHEDID	CHU
BETTI	BOWEN	CONN	COWELL
BIRSS	CABUNALDA	DORAN	DE PLEDGE
BOCCAMAZZO.D	CHARLES	ESTENS	HELSBY
BOCCAMAZZO.C	COLBY	VICKERY	JOHNSON
BOLES-RYAN	CUNNINGHAM		PALAYUKAN
BOTHMA	DIXON		
BOWEN	DORRELL		
CABUNALDA	FORWARD		
CHARLES	GALANTE		
CHU	GESTE		
COLBY	GIANATTI		
COWELL	GORJY		
CUNNINGHAM	HAINES		
DE PLEDGE	HAWKETT		
DIXON	HIGGINSON		
DORRELL	HOLT		
FORWARD	INGRAM		
GALANTE	MASON		
GESTE			
GIANATTI			
GORJY			
HAINES			
HAWKETT			
HELSBY			
HIGGINSON			
HOLT			
INGRAM			
JOHNSON			
MASON			
PALAYUKAN			

Table 9.5: Cohesion Score Grouping results –Score Group "Intermediate"

9.2.1.2.1 Discussion

Showing more promise than the scoring logics for the Audience and Ideas criteria, tests in Precision, Recall and F-measure returned values ranging from 0.68 to 0.69 as shown in Table 9.6, indicating roughly the same performance as for the previous test on the Character and Setting criterion.

With the system placing 12 essays in a better Score Group, it could be said that it was more lenient than the human markers. However, the system also placed 12 essays within this Score Group and 7 out of these 12 were from the higher Score Group, based on human marker assigned scores.

This is leads to the conclusion that the system is neither largely more lenient nor stricter than its human counterparts. This thus leads us to the assumption that the same problems that occurred in the previous test under the Character and Setting criterion are also present here, which is the inconsistency in the human marker scores.

True Positives	False Positives	False Negatives	Precision	Recall	F - Measure
26	11	12	0.69	0.68	0.68

Table 9.6: Precision, Recall and F-Measure results for Cohesion – Score Group "Intermediate"

Even with the majority of the essays sorted into this group, the scoring logic still managed to perform reasonably well. According to Table 9.7, for 25 essays agreed upon out of the 37 placed in this group by the human markers, the system returned an exact agreement rate of 67%.

Score Group	System	Human Makers	Exact Agreement Rate
Intermediate	25	37	67.57%

Table 9.7: Exact Agreement Rate for Cohesion – Score Group "Intermediate"

9.2.1.3 Score Group – Good

Containing the smallest portion of the test essays, this Score Group consisted of 19 essays when sorted according to their human marker assigned band scores. Of these

19, the system identified 12 essays that belonged to the same Score Group, while disagreeing on 7. Table 9.8 shows the details.

Actual	System (True Positives)	Errors	
		False Positives	False Negatives
BELLIS	BELIIS	AMESS	BOTH-WATSON
BOTH-WATSON	BRAMPTON	BAKER.L	CASTAING
BRAMPTON	BREAN	BIRSS	CATOVIC
BREAN	BYRNES	BOCCAMAZZO.D	CHANDLER
BYRNES	COPPARD	BOLES-RYAN	CHEDID
CASTAING	DANKS	BOTHMA	DORAN
CATOVIC	FOO	CHU	VICKERY
CHANDLER	KARSKI	COWELL	
CHEDID	KROLL	DE PLEDGE	
COPPARD	LOH	HELSBY	
DANKS	MAIN	JOHNSON	
DORAN	PERSSON	PALAYUKAN	
FOO			
KARSKI			
KROLL			
LOH			
MAIN			
PERSSON			
VICKERY			

Table 9.8: Cohesion Score Grouping results – Score Group "Good"

9.2.1.3.1 Discussion

Compared to the previous test regarding the same Score Group for the Character and Setting criterion, the scoring logic here performed slightly better, although the Precision is still found rather wanting. With the number of false positives equalling the number of true positives, the Precision score came to only 0.50, as shown in Table 9.9. The 7 essays that the system disagreed on were instead found one Score Group lower. In addition, the 12 false positives that occurred all belonged to the "Intermediate" Score Group. The Recall and F-measure scores fared slightly better, with scores of 0.63 and 0.56 respectively.

True Positives	False Positives	False Negatives	Precision	Recall	F - Measure
12	12	7	0.50	0.63	0.56

Table 9.9: Precision, Recall and F-Measure results for Cohesion – Score Group "Good"

Once again it seemed that the scoring logic was not as accurate when determining whether an essay should belong to this Score Group when compared with the groupings based on their human marker scores. Although this stage of the evaluation returned poorer results, the same could not be said for the exact agreement rate.

As described in Table 9.10, with the system agreeing on 12 of the 19 essays that, based on the human marker scores were placed in this Score Group, the exact agreement rate came to 63%, which was much higher than the previous test's result from this scoring logic of only 38%.

Score Group	System	Human Makers	Exact Agreement Rate
Good	12	19	63.16%

Table 9.10: Exact Agreement Rate for Cohesion – Score Group "Good"

9.3 Methodology Stage 2 – Hypotheses Testing

Based on the features described in the conceptual framework of group scoring essays under the Cohesion criterion, the following hypotheses were generated for this part of the thesis:

The null hypothesis, $\mathbf{H_0}$ is described as:

> $\mathbf{H_0}$ **:** There is no significant difference between the human marker scores and the machine-generated scores under the Cohesion criterion

The alternate hypothesis $\mathbf{H_1}$ would thus be:

$\mathbf{H_1}$: There is a significant difference between the human marker scores and the machine-generated scores for the Cohesion criterion.

9.3.1 Chi-Squared Goodness of Fit Test Results

Group	Observed	Expected
Poor	31	34
Intermediate	37	37
Good	24	19

Based on the data, the formula for determining the chi square was:

$$x^2 = \frac{(31-34)^2}{34} + \frac{(37-37)^2}{37} + \frac{(24-19)^2}{19}$$

$$x^2 = 0.26 + 0 + 1.31$$

$$x^2 = 1.58$$

DF	P=0.995	P=0.975	P=0.9	P=0.5	P=0.1	P=0.05	P=0.05	P=0.01	P=0.005
1	0.000	0.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838

As stated previously, for the null hypothesis to be rejected, the value of P should exceed 0.95. The P value obtained through a comparison to the highlighted row (DF 2)

on Chi-square distribution is between 0.1 and 0.05, which is insufficient to reject the null hypothesis.

9.3.2 Paired T-Test Results

$$t = \frac{0.08-0}{\sqrt{0.48/(89)}}$$

$$t = 1.71$$

DF	0.1	0.05	0.02	0.01	0.005	0.002	0.001
89	1.6622	1.9870	2.3690	2.6322	2.8787	3.1844	3.4032

For there to be a significant difference between the machine and human marker scores, the value of t should exceed the t critical value which is at 1.98. With the t value at 1.71, it is thus shown that there is insufficient evidence to suggest a great dissimilarity between the machine and the human marker scores.

Hence, the null hypothesis $\mathbf{H_0}$ is kept since it cannot be rejected leading us to reject the alternative.

9.4 Conclusion

This chapter discussed the results gathered from the tests carried out to determine the accuracy of the system when using the scores assigned by a human marker as the basis for evaluation. The findings indicate that the scoring logic is extremely proficient in identifying essays that are of a poor quality while having mixed success with the other

two Score Groups. Tables 9.11 and 9.12 show the average values for the first and second stages of the evaluations respectively.

If we were to take the first two tests as any kind of precedence, the expected performance for the "Intermediate" Score Group section should have been much poorer, although the results have shown otherwise. It should be noted, however, that even though results were better than expected, the scoring logic for this particular Score Group is by no means perfect, and has much room for improvement.

Overall, the system seems at times to be more lenient, with 15 essays grouped higher than was done by the human markers but at other times stricter, placing 7 essays in a lower Score Group. On average, the system managed to achieve reasonable performance results for the first stage, with scores of 0.73 across the performance metrics of Precision, Recall and F-measure.

	Poor	Intermediate	Good	Average
Precision	1	0.69	0.50	0.73
Recall	0.88	0.68	0.63	0.73
F- Measure	0.94	0.68	0.56	0.73

Table 9.11: Precision, Recall and F-Measure results for Cohesion criterion

In terms of exact agreement rates, the system showed more promising results, with rates ranging from 63% to 88%, averaging at a 74% exact agreement rate.

Score Group	System	Human Makers	Exact Agreement Rate

Poor	30	34	88.21%
Intermediate	25	37	67.57%
Good	12	19	63.16%
Overall	69	90	74.44%

Table 9.12: Exact Agreement Rate for Cohesion criterion

Even though there seems to be some disparity when comparing the scoring process of the system with that of its human counterparts, the average values of 0.73 in F-measure and 74% in exact agreement rate do surpass the minimum threshold. This allows us to conclude that the features identified earlier in this chapter, with emphasis on the type of connectives found, are related to the grade that an essay receives for the Cohesion criterion. Based on the results gathered from the Chi square and t-tests, it was concluded that there was insufficient evidence to reject the null hypothesis$\mathbf{H_0}$, therefore the alternative is rejected.

This end of this chapter concludes the experiments section of this thesis.

Chapter 10 – System Evaluation

10.1 Introduction

In Chapters 6 to 9 we examined the performance of the scoring process for each of the four criteria focused on in this thesis. The objective of this chapter is twofold: the first is to illustrate how the scores from these four are combined to produce an overall score for an essay and the second is to determine the agreement rate in the overall scores between the system and human markers. Since the work done in this thesis covers only four of the ten criteria of the NAPLAN rubric, to obtain an overall score, we take the scores of an essay obtained for the other criteria as assigned by the human marker to make up the total score. Prior to this, let us briefly recap the basic processes which are carried out to automatically determine an essay's score for each of the four criteria.

Firstly, the student essays are put through the Text Analysis Stage, using the Part of Speech Tagger and customised Named Entity Recognition tool to process the text. This allows the system to produce outputs in the form of POS tags and named entities, which are then used in the Score Grouping stage, made up of two parts: the Event Detection and Rubric Formalisation phases.

The Event Detection phase uses the tags and named entities to determine whether or not a sentence qualifies as an Event by looking for three things:

- an Actor,

- Action, and
- State

If a sentence contains all three properties, it is classified as an Event. This is done for all sentences in the essay.

The Rubric Formalisation phase is itself a two-step process. Firstly, the features which correspond to what is perceived to be sufficient in order to achieve a particular band score are determined. These features are then weighed according to their significance in relation to that particular criterion.

The second step takes into account the certain specified conditions that place an essay in one of the three categories namely A, B or C, with C being the poorest. Once these steps have been performed, the outputs from both these processes are combined and used to determine the score group to which an essay belongs.

Once all the above steps have been completed, we are able to assign a score to the essay. Earlier in this thesis, it was mentioned that each criterion has a different number of band scores and the higher band scores are often difficult to tell apart. Therefore, a fuzzy representation was used to allow for a more uniform grouping. In cases where more than one band score are grouped together, a median is assigned. For example, for the Ideas criterion where the Poor grouping consists of band scores 1 and 2, a score of 1.5 is assigned. Table 10.0 illustrates this further.

		Scores assigned to Score Groups		
		Poor	*Intermediate*	*Good*
Marking Criteria	*Audience*	2	4	5.5
	Ideas	1.5	3	4.5
	Character & Setting	1.5	3	4
	Cohesion	1.5	3	4

Table 2.1: Sores assigned to Score Groups

10.2 Assigning the Scores

The following sections provide examples of how the scores assigned by the system are combined to give an overall score. A total of eight essays are examined in this section, which consist of a mix of low, middle and high scoring essays.

For each example, an excerpt of the essay is shown, followed by its score for the respective criteria. Finally, the marks assigned by the system and those by human markers are compared and discussed. For each comparison Table, the criteria for which the system automatically assigns a mark are highlighted in red.

10.2.1 Essay 1

the trser on cutles reef

awitethightsgostly winds in the port of Port Yole. out on Pirte Ship emeges from the bug to steel the mup to the treser on cutles reef and scull sand. and all of a Sudden boom. Thay shoot ther cannons. And ther Best Prite Brent Selversowrd. Sneeks into the moors canter and steels the map. Just then a gard finds him aa fight begins

Brent is victorys. Lets get out of this place. Thay set sail for scull island. But what thaydont know is the monsters on the island who defend the treser. Captain yes Brent i got it good whe will Be ther in 1 month.

Criteria Grouping:

- Audience - Poor
- Ideas - Poor
- Character and Setting - Poor
- Cohesion – Poor

Criteria	Audience	Ideas	Char.& Setting	Cohesion	Text Structure	Vocab.	Para.	Sent. Structure	Punct.	Spelling	Overall
Human Marker	2	2	2	2	2	3	1	2	2	3	21
System	2	1.5	1.5	1.5							19.5

10.2.2 Essay 2

AaaahhhhEdmandwathabend. “I vell into a hole on a dead boddy”

“Try to call the police”, Josh please Im scared I have no signel on my sellfone. I will go on get som help Edmand.

I rame into so villigers. “I nicely told wat happened.” They canotinderstand. I grabed a

rope an run speedily to the hole I droped the rope into the hole to help my friend.

He grabd the rope an I dragde him and the man who is dead.

We drage the boddy to the villigers

I said “We found dis boddy in a huge hole.” “Do you understand.”

Bihind me someone said yes.

It’s “Maria jelled ‘Josh how we lost with our last trek at Treaser Island.

Criteria Grouping:

- Audience - Poor
- Ideas - Poor
- Character and Setting - Poor
- Cohesion - Poor

Criteria	**Audience**	**Ideas**	**Character & Setting**	**Cohesion**	**Text Structure**	**Vocab.**	**Para.**	**Sentence Structure**	**Punct.**	**Spelling**	**Overall**
Human Marker	2	2	1	2	2	2	1	2	2	2	18
System	2	3	1.5	1.5							19

10.2.3 Essay 3

One day i was walking in the park and i found a dog. Nobody was with the dog at the time. i walked to back home with the dog and i told my mum + dad that i found a dog at the park. The dog was only a puppy and he did not now were his mum went that day. the next day i found the dog's mum. The dog mum was at the vet and. She was asleep. In a few weeks timei meet a friend and i found something out about her. It was that. The dog i found it was her dog. that ran away i did not know that it was her dog she stayd at my house 1 night and we keep the dog.

Criteria Grouping:

- Audience - Poor
- Ideas - Poor
- Character and Setting - Poor
- Cohesion - Poor

Criteria	Audience	Ideas	Character & Setting	Cohesion	Text Structure	Vocab.	Para.	Sentence Structure	Punct.	Spelling	Overall
Human Marker	2	2	1	2	1	2	0	2	2	3	17
System	2	1.5	1.5	1.5							16.5

10.2.4 Essay 4

"Dad," "yes"! "can I sleep at my friends"? "sure!" "so were you going" said mum "to my friends, is that ok? saidsean "Well Ok but be good." said mum "I will" said sean. "I will

take you” said mum “bye be good good.” said dad. So Di takes sean to his friends. “Im home” said mum “wow that was fast”

* 4 hours later * “knock knock” said the officer

Criteria Grouping:

- Audience - Poor
- Ideas - Poor
- Character and Setting - Poor
- Cohesion - Poor

Criteria	Audience	Ideas	Character & Setting	Cohesion	Text Structure	Vocab.	Para.	Sentence Structure	Punct.	Spelling	Overall
Human Marker	2	2	1	2	1	2	1	2	1	2	16
System	2	1.5	1.5	1.5							15.5

10.2.5 Essay 5

Jessica Starlett was an ordinary A.S.H.S. student. She was 14 years old and was about 160cm tall. She had Blonde hair that went down to her shoulder blades and had brown eyes. One day, witch seemed to be any ordinary day. She was in the canteen with her friends having some lunch, when this girl approached them. Jessica almost gasped at the site of this girl. The was very pale, had black eyes with black ‘sleep’ rings around, and had hark, brown, ragged hair. She went up to Jessica and said, “My name is Natalie

Fisher and the mirror ghost has asked me to deliver this letter to you…..” she gave the letter to Jessica. Jessica read the letter in her head, it said

Criteria Grouping:

- Audience - Intermediate
- Ideas - Intermediate
- Character and Setting - Intermediate
- Cohesion - Intermediate

Criteria	Audience	Ideas	Character & Setting	Cohesion	Text Structure	Vocab.	Para.	Sentence Structure	Punct.	Spelling	Overall
Human Marker	4	3	2	3	3	3	2	3	3	3	29
System	4	3	3	3							30

10.2.6 Essay 6

Her raven hair glistened in the sunlight as she flowed past me, leaving a scent of lavender in the air. She looked back at me, laughing, and said, “Come on. Elisa! You’re so slow!” I ran up to her, laughing back, my heart full of the joy that seemed to radiate out of her.

She was flowing further away, my legs stopped moving, I called out to her, “Come back! Mum come back!” But she slipped further and further away, fading as she went, A few seconds later, she was all but gone…

I jerked my eyes open, breathing rapidly. A shiver of fear shot down my spine,

before a brilliant light shone into my eyes, burning my retinas I clamped my eyes shut and groaned. What a dream, I mused. Then, with another jolt down my spine, I remembered the date toady. It was my birthday. But with that realisation, I remembered what today also meant. No, I told myself. Today was not the day to dwell on that.

Criteria Grouping:

- Audience - Good
- Ideas – Good
- Character and Setting - Good
- Cohesion - Good

Criteria	**Audience**	**Ideas**	**Character & Setting**	**Cohesion**	**Text Structure**	**Vocab.**	**Para.**	**Sentence Structure**	**Punct.**	**Spelling**	**Overall**
Human Marker Score	6	5	4	4	4	5	2	6	4	5	45
System Score	5.5	4.5	4	4							44

10.2.7 Essay 7

" Comeonnn!! Pleeasee do my homework for me! I really hate maths – and besides, what's a few favours between friends right?"

Joanna sighed as she watched the nameless girl being sucked into the trap of that

wheedling blonde. This wasn't the first time, either. The nameless girl had been coerced into doing everything for the blonde – Kaitlyn, her name was – from homework to household chores, even so far as polishing shoes! All in the name of friendship. 'That doesn't give much of a good reputation to friendship huh?' Joanna thought.

Criteria Grouping:

- Audience - Good
- Ideas - Good
- Character and Setting - Good
- Cohesion - Good

Criteria	**Audience**	**Ideas**	**Character & Setting**	**Cohesion**	**Text Structure**	**Vocab.**	**Para.**	**Sentence Structure**	**Punct.**	**Spelling**	**Overall**
Human Marker Score	6	5	4	4	4	4	2	6	5	5	45
System Score	5.5	4.5	4	4							44

10.2.8 Essay 8

On the highway, all walking, all following, all bustling to get to the front, to not be left behind. I jostle along with the others, trying to be important but still not straying from the crowd. This is the network of roads and paths that is life.

Finally sick of the infinite push and shove of this life, I turn off onto a narrow road, scattered with people, randomly and sparsely. I follow this road, and then turn out a single-lane, narrow and overgrown street. Only a few people wander along here, scattered few and far between. It is pretty here, in a sort of solitary silence. Neglected and almost forgotten, those who walk this way like the unique and individual things in life, and have a love for beauty.

Criteria Grouping:

- Audience - Intermediate
- Ideas - Intermediate
- Character and Setting - Good
- Cohesion - Intermediate

Criteria	Audience	Ideas	Character & Setting	Cohesion	Text Structure	Vocab.	Para.	Sentence Structure	Punct.	Spelling	Overall
Human Marker Score	6	5	4	4	4	5	1	6	5	6	46
System Score	5.5	4.5	3	4							44

This section presented excerpts of eight sample essays selected at random. A brief evaluation of the scores indicates that the system appears to be slightly stricter than the human markers. Also, it is known that due to the moderation of the scores for each criterion (the maximum score the system would give for the Audience and Ideas criterion are 5.5 and 4.5 respectively), the essays are graded according to a maximum of 45.5 instead of 47.

While this difference in scale may cause some errors in scoring, from the examples above, the scores differ only by a small margin of 1-2 points. If this is indeed the case for most of the essays graded, then it stands to reason that the slight difference in scale would not have a significant impact on the overall grading of the essay. That is, a good essay is still scored as such and there should be no instances of a good essay scoring poorly. The next sections will further investigate the above statements in the overall evaluation of the system.

10.3 Evaluation - System Scores vs. Human Scores

The section above illustrated the method by which the system-assigned scores were obtained and combined. It was also proposed that even though the system scores the essays on a slightly modified scale, this would not have a large impact on the overall score. In this section, we aim to determine if that is indeed the case; at the same time, we evaluate the system's overall agreement rate with regards to the final scores in comparison with those of the human markers.

10.3.1 Distribution of low to high scoring essays

In the first step taken to determine how similar the system scoring is to that of the human markers, the distribution of low to high scoring essays is taken into account.

While this is does not give a specific value of agreement, it allows us to see if there is a large difference in the scoring trends between the two. If, for example, the system scores a significantly higher number of essays in a manner contrary to the scoring by the human markers, it can be assumed that even without further testing the system would not be a viable alternative. Conversely, should there be a similar grouping of the scores, it could be said that the scoring trends of both the system and human markers are relatively the same. According to Stemler (2004), such a measurement provides some insight into common scoring trends rather than being a measurement of error itself.

For this purpose, the essays were split into four groups, namely essays that score between:

- 1-20
- 21-30
- 31-49 and;
- 40 and above

The results obtained are shown in Figure 10.1.

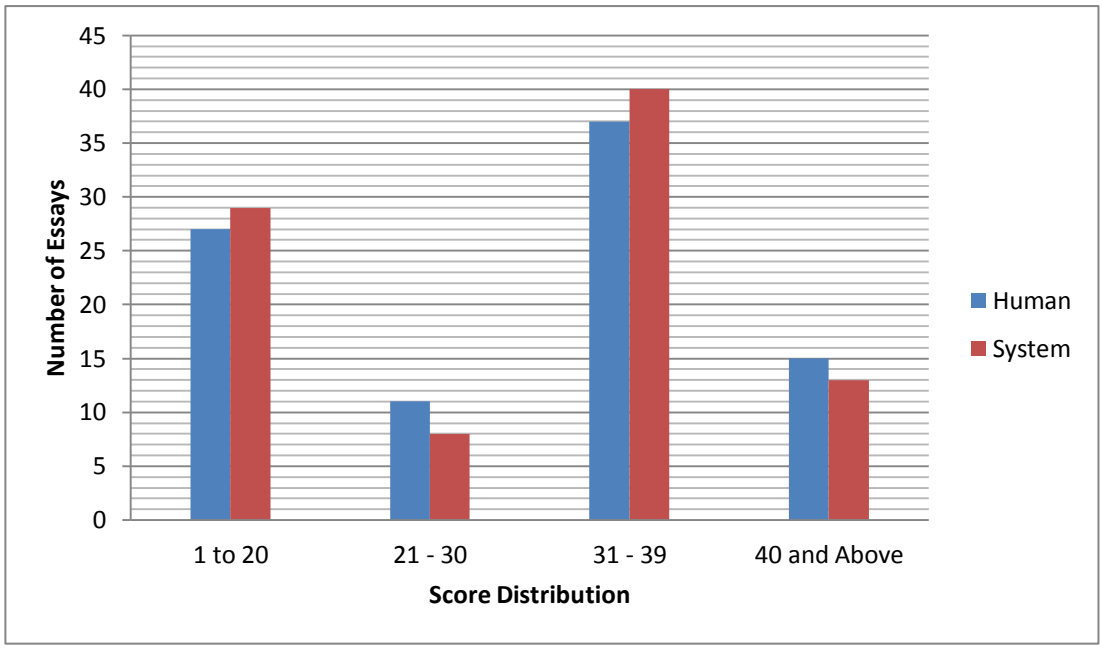

Figure 10.1: Score Distribution

As can be seen in Figure 10.1, the number of essays in each group is more or less the same, with no more than a difference of three essays within each respective group. Thus, it is assumed that the system and human markers are generally similar when scoring an essay. For the purpose of further analysis, each individual essay score assigned by the system was compared with that given by the human markers. The results of this are shown in Figure 10.2.

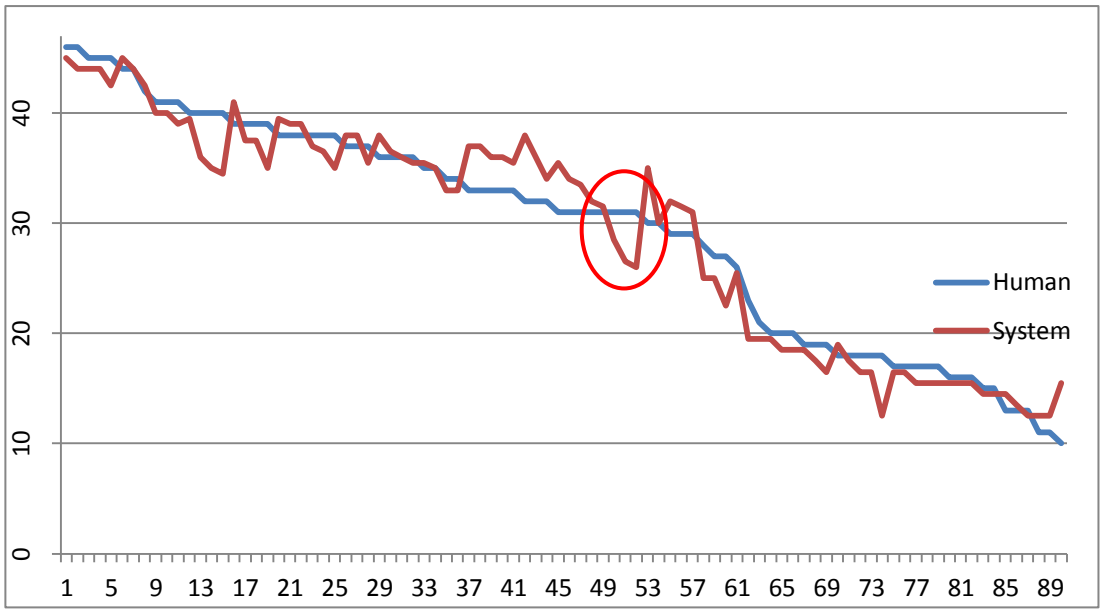

Figure 10.2: Comparison of Individual essay scores (Human vs. System)

As expected, the results in Figure. 10.2 coincide with those shown in Figure. 10.1, although a potential problem, circled in red, is highlighted. These essays are cause for some concern not only due to a larger difference in marks awarded, but also because this difference has caused them to be placed in a different grouping.

However, in the overall scheme of things, there seems to a significant similarity in the scoring trends between the system and human markers. This allows us to proceed with further evaluations of the specific level of agreement between the two. For this purpose, the adjacent agreement rate is discussed.

10.3.2 Adjacent Agreement Rate

The use of exact agreement rates, where the system and human marker scores are exactly the same, does present a more accurate representation of agreement which is difficult to achieve even with a short scoring scale. Therefore, consensus

measurements such as the adjacent agreement rate allow for a more robust analysis (Brown 2004).

In most cases, the adjacent agreement rate depicts the percentage of where the system and human marker scores differ by only one point (Larkey 1998). However, in a larger point scale it might be extended slightly using the following equation:

$$\text{Adjacent Agreement} = \%(\, |\text{Truescore} - \text{Systemscore}| < a)$$

Formula 10: Adjacent Agreement

Where a = the maximum accepted difference between the two scores.

It is important to note that the scale has to be sufficiently long in order to offset the instances where the system and human markers agree through chance. In the case of this thesis, the 47-point scale is rather large and would thus account for random chance agreements.

In recent studies, adjacent agreement rates usually vary between rates of 80-100% (Brown 2004). Therefore, as a minimum standard, the adjacent agreement rate for the system should not fall below 80%.

Furthermore, in the previous section, as an additional condition the clustering of high and low scores should be relatively similar in order to justify a larger allowable difference in scores. Figure 10.1 showed that the clustering was indeed relatively similar and from the samples shown earlier, the grade of the essay would not be greatly affected if the difference is less than or equal to 3, thus setting the threshold for a in the above equation.

Figure 10.3 below illustrates the full range of differences between the scores given by the system and the human markers. For the full comparison of the scores given by the system and the human makers, refer to Appendix I.

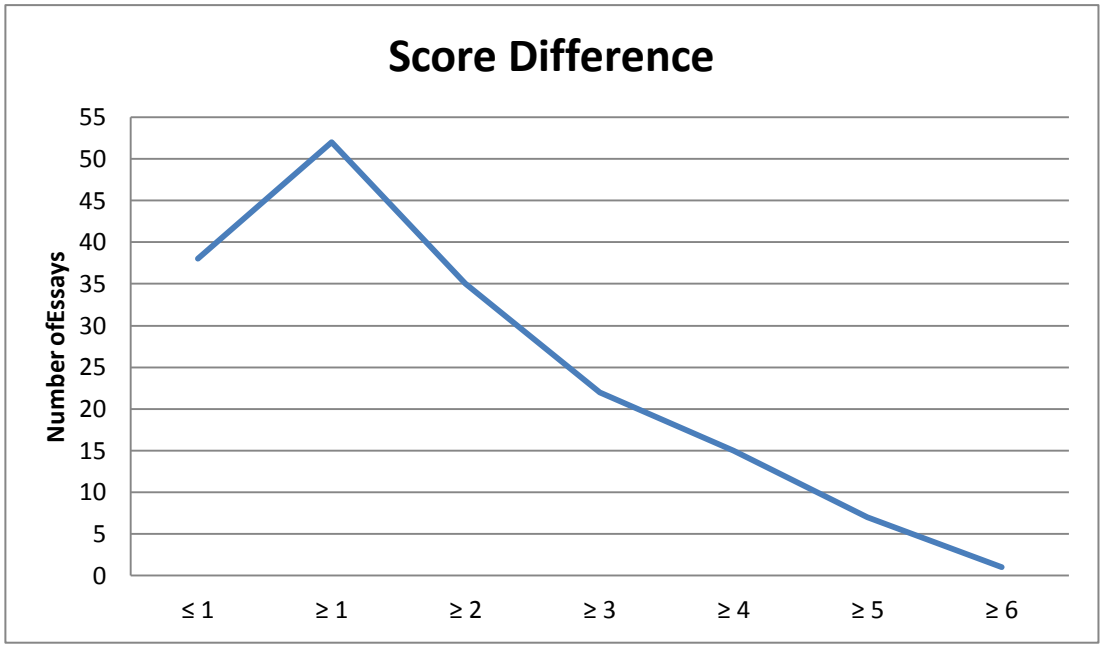

Figure 10.3: Score difference between system and human markers

Therefore, the adjacent agreement rate for the system can be represented as:

$$\text{Adjacent Agreement Rate} = \%(\ |\text{Truescore} - \text{Systemscore}| \geq 3)$$

Based on the results, there was a difference of more than 3 points in 16 essays out of a total of 90, therefore:

$$\text{Adjacent Agreement Rate} = 74/90 = 82\%$$

Based on the number of times the scores differ by no more than 3 points, the system achieves an adjacent agreement rate of 82%, which allows us to conclude that the system is in sufficient agreement with the human markers.

10.4 Conclusion

This chapter provided a brief review of the steps taken in order to automatically acquire an essay's overall score using the output gathered from analysing the four criteria which were the focus of this thesis. Since there are a total of ten criteria in the NAPLAN marking rubric on which this thesis is based, the scores for the remaining six are taken from the human markers in order to obtain an overall score.

A total of eight essay excerpts were selected of the test bed of 90 and had their system-assigned scores compared with the original human marker scores. The score differences for all the samples were below 3 points.

However, in order to determine if the system was indeed sufficiently in agreement with the scores given by the human markers, several evaluation measures were applied.

Firstly, the system had a scoring trend similar to that of the human markers; this allowed us to conclude that the two were in agreement to a degree and that the small difference in the marks given would have little effect on the eventual overall grade of an essay.

Secondly, the adjacent agreement rate was examined. Having established earlier that a small difference of up to 3 points would have little impact on the eventual grade of an essay, the threshold for allowable difference was therefore extended, which gave the system an adjacent agreement rating of 82%.

With a similar scoring trend and an adjacent agreement rate of 82%, it was thus concluded that the system was in sufficient agreement with the human markers and thus could serve as a viable approach to automated essay grading.

The next chapter brings this thesis to its conclusion, recapitulating the work done and discussing the challenges encountered during the course of this work.

Chapter 11- Recapitulation and Future Work

This thesis has described an essay scoring method that combines concepts from Narratology and Automated Essay Grading technology, in an attempt to provide a novel method of essay grading. This goal was achieved in part with the Event Detection framework, which was able to effectively identify within an essay those sentences that were integral to the story line. Based on this output, an essay grading system was developed which was able to determine with relative accuracy an essay's score based on four criteria from the NAPLAN marking rubric.

This chapter summarizes the work undertaken in the development of this system and discusses some avenues for future works.

11.1 Recapitulation

The development of AEG systems was always on the horizon given the multitude of additional duties teachers have to undertake.

In Chapter 1, a preliminary insight into the fields of Automated Essay Grading was provided, giving some of the main reasons why these systems were developed and how they have affected us thus far. In addition to providing an introduction to some of the challenges faced by automated systems, some past and on-going debates regarding the use of AEG systems were presented in which both sides of the argument were examined briefly.

Apart from introducing AEG technology, this chapter also examined briefly the field of Narrative Analysis. From this field, it was anticipated that this thesis would extract various concepts regarding free text narrative type essays and attempt to grade an essay using these concepts as a foundation.

In Chapter 2, a literature review was conducted to examine the previous works within the fields that relate to Automated Essay Grading. Through a review of the literature, several methods by which essays are graded were identified. These ranged from systems that utilise the more technical statistical methods of linear regression and vector space computations to the ones that seek to identify the more subtle aspects of free text answers using Natural Language Processing tools, Artificial Intelligence and Neural Networks.

Each system has its fair share of advantages when it comes to essay grading; statistical and prediction methods were able to perform fairly well but lacked the ability to process text on a more contextual level; while other methods, using NLP tools, AI technology and the like were able to identify the more implicit features of free text responses but required more computational power and possibly more human supervision than did their statistical-based counterparts.

Mainly, the issues and challenges that plague AEG systems were that those adopting the statistical approach needed a way to deal with more implicit features of free text, while other systems which did manage this were highly dependent on the type of student response (short answers or responses where keywords were looked for), the

system having pre-knowledge pertaining to the subject domain in addition to requiring more computational power. It was also found that while narrative essays are one of the main methods by which a student might be able to showcase his/her mastery of the English language, no system thus far has attempted to incorporate narrative concepts into the analysis of these types of free text responses.

Research questions derived from the literature review became the basis for Chapter 3, in which the problems to be addressed in this thesis were identified and discussed.

Chapter 3 formally defined the research issues that are the focus of this thesis. In addition, the key concepts that are used throughout this work were explained in detail. As mentioned in Chapter 2, it was established that the main challenges faced by AEG systems are that they are sometimes unable to process text on a more contextual level; they require a large amount of resources (training data, computational power, etc.); or are highly dependent on the subject domain. Therefore, the research aims that were stated in this chapter were intended to address these issues.

A discussion of the various methodologies by which these aims could be achieved was also provided, which led to the choice of a hybrid methodology which applies concepts of both the Science and Engineering and Social Science methodologies.

In Chapter 4, the various narrative analysis concepts that influenced the formulation of the proposed solution were discussed. It was decided that the proposed solution would be centred on the detection of Events. However, since simply detecting these Events

was not sufficient to generate a grade which related to each of the four criteria, there needed to be a way to determine how the Events themselves are related to the criteria.

This was done in the Score Grouping stage of the solution which identified those features of an essay that most significantly impact on the grade an essay receives for the respective criteria of Audience, Ideas, Character and Setting and lastly Cohesion. By mapping certain features of an essay and focusing on the detection of Events, it is possible to grade an essay according to these criteria.

Before any of the above analysis stages could be carried out, the raw text needed to be processed into machine-readable output. This was done in the Text Analysis Stage where Natural Language Processing tools were applied to the raw text to transform it into the desired machine-readable output. This output was then used in the Score Grouping stage which was made up of the Event Detection and Rubric formalisation processes. Following that, the main Natural Language processing tools used in processing the raw text were presented together with an overview of the solution designed to tackle those issues and address the stated aims.

The theoretical framework for the proposed solution was explained, detailing the Text Analysis stage in which the student responses are pre-processed into various output types. These outputs were in turn used as inputs for the Event Detection Stage and eventually the Rubric Formalisation Stage in which essay scores were determined based on the criteria from the NAPLAN marking rubric.

Having shown how the concept of an Event from the context of narrative texts is extracted, the Rubric Formalisation Stage also detailed how these concepts formed the basis of the scoring processes of the Audience, Ideas, Character and Setting and Cohesion criteria.

Chapter 5 described in finer detail the processes that made up the Event Detection Stage and its performance. The specific text processing steps, mainly involving Named Entity Recognition and Part of Speech tagging used for the desired format for further processing, were explained.

After describing how a sentence would be classified as an Event and explaining how these Events are detected within a narrative essay, the method was tested using a bed of 1340 sentences, previously annotated manually according to whether or not each was considered as an Event. Of the 682 sentences classified by human markers as an Event, 658 sentences were correctly classified as such by the system. In additional tests to evaluate the system's accuracy, it achieved an average Precision, Recall and F-measure score of 0.85, 0.85 and 0.84 respectively. Taking into account the true negative rate, the system gave an average Matthew's Correlation Coefficient score of 0.52. Overall, the Event Detection process performed at an acceptable level for its output to be used in the next stage of the proposed solution.

Chapters 6 to 9 evaluated the performance of the scoring process according to the NAPLAN rubric. This evaluation was done in two stages. The first utilised several performance measures such as Precision, Recall, F-measure and exact agreement rate.

The accuracy of the system in correctly grading an essay according to the human marker scores is represented in whole using the F-measure, while the exact agreement rate was a measure of the similarities between the system and human markers when assigning an essay to its relevant Score Group.

In the second stage, the hypotheses were tested to determine whether there was indeed a similarity between the grades assigned by the computer and those given by the human markers. The two statistical methods used were the Chi-squared and paired T-test. The Chi-squared test shows how close the values of the observed data are to those of the expected values while the paired T-test showed us whether there was a significant difference between the means of the scores generated by the computer and those of the human markers. In order for the null hypothesis to be rejected, thereby indicating a significant difference between the two, the Chi-squared and T-test value had to exceed 7.3 and 1.98 respectively.

Chapter 6 dealt with the Audience criterion of the NAPLAN rubric. The features considered for this criterion were the essay's length, the number of Events, the Event Ratio and the presence of a Physical or Mental State. In terms of accuracy, the scoring logic attained an average F-measure score of 0.72 while achieving an overall exact agreement rate of 75%.The Chi-squared and paired T-test value came to 1.93 and 0.52 respectively, which was insufficient evidence to reject the null hypothesis that there was no significant difference between the system and the human markers.

Chapter 7 focused on the Ideas criterion, using features such as the number of unique adjectives and adverbs, in addition to the base features such as the number of Events, Event Ratio and Essay Length. Results from the experiments gave an average Precision, Recall and F-measure value of 0.67 while showing an exact agreement rate of 70%. The Chi-squared and paired T-test value came to 0.44 and 0.72 respectively, which was again insufficient evidence to reject the null hypothesis that there was no significant difference between the system and the human markers.

In Chapter 8, the performance of the system for the Character and Setting criterion was explored. In addition to the aforementioned base features used, the scoring logic for this criterion included features such as the number of unique adverbs and adjectives, together with the presence of a physical or mental State. The scoring logic returned an average value of 0.62 across Precision, Recall and F-measure while giving an exact agreement rate of 71%.The Chi-squared and paired T-test value came to 0.98 and 0.48 respectively, indicating that there was insufficient evidence to reject the null hypothesis that there was no significant difference between the system and the human markers.

Chapter 9 detailed the experiments conducted on the last criterion, Cohesion. Unique features which make up the scoring logic included the number of simple and advanced connectives, checked through the use of a simple lexicon of connectives. The resulting scores when measuring Precision, Recall and F-measure values all returned an average of 0.73, while returning an exact agreement rate of 74%.The Chi-squared and paired T-test value came to 1.58 and 1.71 respectively, indicating that there was insufficient

evidence to reject the null hypothesis that there was no significant difference between the system and the human markers.

The results gathered from experiments conducted on the four criteria mentioned above indicate that the means of detecting Events within a narrative type story, when applied to essay grading, does impact on an essay's final score. All experiments achieved an average F-measure score of 0.65 and above while exact agreement rates were no lower than 70%. Chi-squared and paired T-test values all indicated that there was insufficient evidence to show that there was a significant difference between the scores generated by the computer and those of the human markers.

Chapter 10 illustrated how scores could be combined based on the results gathered in Chapters 6-9. In addition, the scoring trends of the system and human markers were found to be relatively similar, while the adjacent agreement rate between the system and human markers was found to be 82%.

11.2 Contributions

This research has shown that through the use of simple text mining and NLP techniques, it is possible to detect what this thesis previously defined as an Event, which can be seen as the more important parts of a story within a Narrative context. The other major contributions of this thesis are as follows:

11.2.1 Novel Method of Essay Grading

The application of Narrative analysis in the field of Automated Essay Grading has also opened up a new direction of analysis for future researchers. In addition to providing a novel method of essay grading, the system requires neither heavy computation nor pre-knowledge of a subject domain, which would mean that potential costs involved in implementing this system would be predictably low.

11.2.2 Independent of subject domain

Most AEG systems developed so far deal with student responses to prompts from a specific subject domain. While constrained by the rubric template used for essay assessment, the grading system itself is entirely independent of the subject domain. This means that whichever marking rubric is used (narrative or persuasive writing) the subject matter does not affect the grading process. This is highly advantageous since the system need only be trained once on that particular rubric, and not repeatedly according to the subject domain.

11.2.3 Scoring model only needs to be trained once per writing genre

As mentioned earlier, since there could be an unlimited number of subject domains for which a student might be asked to write a narrative essay, the creation of an essay grading system that needs to have pre-knowledge of a particular domain would be unfeasible. Therefore, due to the domain-independent nature of this grading system, there need be only one training stage per writing genre (narrative, persuasive, etc.) and

the system would then be able to automatically grade essays according to the features of that particular genre.

11.3 Challenges

There are some inaccuracies that will inevitably accompany the method of detecting Actions since words such as ‘was’, ‘do’ or ‘can’ are verbs. However, the inaccuracies would have a minimal effect on whether or not a sentence is considered as an Event. Most of the time, these verbs that do not constitute an Action are found within the same sentence as those that do, and so essentially they fulfil the condition for an Action to be present. Furthermore, since a sentence can be classified as an Event only if an Action, Actor and State are all present, and the presence of an Actor is rarely without an Action, the abovementioned problem would have little effect on the eventual outcome.

Other challenges and hurdles that still need to be addressed include:

11.3.1 Dealing with Dialogue

The system would have some difficulty dealing with dialogue between characters. In conversations between characters, there need not be a mention of an Actor since it is already implied that two or more Actors are engaged in the dialogue. This is obvious to the human reader but the system still needs to be able to identify such situations in the narrative. Possible solutions include using the Part of Speech tagger to include tags for quotation marks that might denote the beginning and end of a conversation, thereby

indicating that an Action mentioned within would relate to those Actors, although this is not a totally satisfactory solution.

11.3.2 Spelling Errors

As with other essay grading systems, despite a slew created by spelling errors, the meaning of the text would be fully understood. Having mentioned this, the only time this would really be a problem is when a student has written an excellent essay albeit riddled with spelling errors. Although this is rarely the case (in most cases a student who is able to write a good essay would have minimal spelling errors and mostly in words that serve to add meaning and are not vital to comprehension itself), there still needs to be some measure that accounts for spelling mistakes while not compromising the integrity of the scoring process.

11.3.3 Short Essays

Essays containing a great deal of description would pose a problem since they could have few Actions and many possible States. This might be solved by looking at the number of adverbs and adjectives and determining how many of them are unique. Short scripts of an extremely high quality suffer due to the minimum requirements of the scoring system.

Even though scripts such as these are relatively rare, occurring only twice in 90 essays, they are a cause for concern. While it is possible to predict an essay's grade based on the features characterising a good, intermediate or poor essay, it is obvious that more contextual analysis would be required to improve the system's accuracy. While

weighted features for essay length have been introduced, the system would benefit from a more contextual-based analysis, although this might significantly increase the computational requirements.

11.3.4 Brute Force Methods

Cohesion scoring is largely based on a brute force method of checking against a lexicon of connectives. If there is one word or a group of words that might be seen as a connective but is not in the list, it would be ignored, thereby causing an essay to receive a score lower than it deserves.

11.3.5 System Training

While it is true that the grading system needs to be trained only once, if a new marking rubric were to be introduced, the system would obviously need to be calibrated to match the requirements of the said rubric. However, it is possible to save those calibrations so that several marking rubrics are available that can be applied using the proposed solution.

Further training to detect repetition would also be needed since the system could be tricked should a well-written paragraph be repeated multiple times. Since the system does not conduct contextual analysis past the detection of Events, the number of times the same Event is repeated is not accounted for.

11.4 Future Work

This thesis has concentrated on the rubric for narrative type essays which is available nationally in the NAPLAN rubric. As new rubrics are defined for other categories such as "argumentative" type essays, the criteria for Audience, Ideas, Character and Setting and Cohesion will have to be revisited.

The human marker assigned scores showed great inconsistencies. On several occasions, an essay that was obviously of poorer quality received the same marks as one of a much better quality.

This number of inconsistencies has led to the fault of the system due to incorrectly or inaccurate classification based on the subjectivity of and disagreement between the human markers. However, there is no way to tackle this within the scope of this thesis. For future work, it would be best to ascertain that most human marker assigned scores have a higher agreement rate before they are used to test the system.

Due to the time constraints of this thesis, it was not possible to fully take into account the different clauses that could exist within a narrative story. In this work, only an analysis of Events was conducted which did not take into account how those Events might be linked to one another and/or to which Actor. Future work in this area would lead to a greater cognitive ability of the system since the subjectivity of the Actors could be taken into account. This would probably require the use of a lexicon of terms possibly more comprehensive than the ones used for experiments in this thesis.

The Event Detection method takes into account only the presence of Events and the effect they have on an essay's score. The actual relationship between those Events and specific Actors is not considered here due to the time and resource constraints of this thesis; a deeper analysis would have required a larger lexicon of terms. Therefore, while some relationship has been discovered between the presence of Events and an essay's score, the system is still not able to identify causal connections between those detected Events. This would be the first and main avenue for future work.

It is also noted that the Score Groups of 'Poor', 'Intermediate' and 'Good' are a crisp partition of the possible score range, as such there sometimes seemed to be a slight overlap in the scoring between theses ranges. One way of taking this into account would be to give a fuzzy representation of each of these bands. This would allow us to obtain membership in each of these bands between 0 and 1 and then a Fuzzy Inference Method could be used to ascribe the band scores.

11.5 Conclusion

This chapter has provided a brief summary of the work done in this thesis, from a review of the literature of current AEG systems to how the problems that needed to be addressed were identified and tackled. The work done in this thesis has shown that while it is still difficult to pick out the tacit features of a narrative essay, an automated grading system does not necessarily require large computing powers or overly complex algorithms to achieve its intended purpose.

On the same note, it is still rather ambitious to aspire to create a system that would have the same cognitive abilities as a reasonably intelligent human. While that certainly is the general direction of current research, it is the small steps that bring us closer to that goal that ultimately matters. As one of these steps, this thesis was primarily aimed at adopting concepts found within the field of Narratology, more specifically narrative texts and combining them with essay grading technologies in order to create a new way that a computerised system is able to process the thoughts of the author.

Bibliography

Abbott, H. P. (2002). "The Cambridge introduction to narrative." Cambridge: Cambridge University Press.

Alani, H., S. Kim, et al. (2003). "Automatic ontology-based knowledge extraction from web documents." IEE Intelligent Systems **18**(1): 14-21.

Alfonseca, E. and S. Manandhar (2002). An unsupervised method for general named entity recognition and automated concept discovery Proceedings of the International Conference on General WordNet.

Attali, Y. and J. Burstein (2006). "Automated Essay Scoring with e-rater V.2." Journal of Technology, Learning and Assessment **4**(3).

Bal, M. and E. Tavor (1980). "Notes on Narrative Embedding." Poetics Today **2**(2): 41-59.

Bal, M. (1985). Narratology: Introduction to the Theory of Narrative.

Bal, M. (1997). "Narratology: Introduction to the Theory of Narrative." University of Toronto Press.

Ben-Simon, A. and R. E. Bennett (2007). "Toward More Substantively Meaningful Automated Essay Scoring." Journal of Technology, Learning, and Assessment **6**(1).

Berger, A., L., V. Della Pietra, J., et al. (1996). "A maximum entropy approach to natural language processing." Computational Linguistics **22**(1): 39-71.

Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis.

Bernardo, J. M. (2000). Bayesian theory.

Bikel, D., M., R. Schwartz, et al. (1999). "An algorithm that learns what's in a name." Machine Learning **34**(1-3): 211-231.

Bloom, B. (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook 1, Cognitive Domain. New York;Toronto, Longmans.

Boccono, S. and F. Nack (2005). Supporting the generation of argument structure within video sequences. Proceedings of the sixteenth ACM conference on Hypertext and hypermedia.

Borthwick, A., J. Sterling, et al. (1998). Description of the MENE Named Entity System as used in MUC-7. Proceedings of 7th Message Understanding Conference.

Brill, E. (1992). A simple rule based part of speech tagger. 3rd Conference of Applied Computational Linguistics. Trento, Italy.

Brown, G. T. L., K. Glasswell, et al. (2004). "Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system." Assessing Writing **9**: 105–121.

Burke, K. (1969). A Grammar of Motives, University of California Press.

Burstein, F. and S. Gregor (1999). The Systems Development or Engineering Approach to Research in Information Systems: An Action Research Perspective. The 10th Australasian Conference on Information Systems. School of Communications and Information Management, Victoria University of Wellington, New Zealand.

Burstein, J. (2003). The e-rater scoring engine: Automated Essay Scoring with natural language processing. Automated Essay Scoring: A cross disciplinary approach. M. D. S. a. J. C. Burstein. Mahwah, NJ, Lawrence Erlbaum Associates**:** 113-121.

Burstein, J. and D. Marcu (2000) "Benefits of modularity in an Automated Essay Scoring System." **Volume**, DOI:

Burstein, J., M. Chodorow, et al. (2003). Criterion: Online essay evaluation: an application for automated evaluation of student essays. Proceedings of the 15th Annual Conference on Innovative Applications of Artificial Intelligence, Acapulco, Mexico.

Chambers, N. J., Dan (2008). Unsupervised learning of Narrative Event Chains. Proceedings of ACL-08: HLT, 2008.

Chang, D. S. and K. S. Choi (2005). Causal Relation Extraction using Cue Phrase and Lexical Pair Probabilities. Natural Language Processing . IJCNLP 2004, Springer-Verlag Berlin Heidelberg**:** 61-70.

Chapelle, O., B. Scholkopf, et al. (2006). Semi-supervised Learning, The MIT Press, Cambridge, Massachusetts, London, England.

Cheville, J. (2004). "Automated Scoring Technologies and the Rising Influence of Error." The English Journal **93**(4): 47-52.

Christie, J. (2003). Automated Essay Marking for Content ~ Does it Work? Seventh Internation Computer Assisted Assessment Conference, Loughborough University, Leicestershire, UK.

Chung, G. K. W. K. and J. H. F. O'Neil (1997). Methodological Approaches to Online Scoring of Essays. CSE Techical Report 461, University of California.

Collins, M. and Y. Singer (1999). Unsupervised Models for Named Entity Classification. Proceedings of the Joint SIGDAT Conference on Empirical Models in Natural Language Processing and Very Large Corpora.

Conlan, G. (1986). Objectives measures of writing ability. New York, Longman**:** 109-125.

Cotos, E. and N. Pendar (2008). Automated Diagnostic Writing Tests: Why? How?, Ames, IA: Iowa State University.

Creswell, J. W. (1998). Qualitative Inquiry and Research Design: Choosing Among the Five Traditions, Thousand Oaks Ca: Sage Publications.

Cutting, D., J. Kupiec, et al. (1992). A practical part of speech tagger. 3rd Conference on Applied Natural Processing, ACL.

Daille, B. (1994). "Study and implementation of combined techniques for automated extraction of terminology." The Balancing Act: Combining Symbolic and Statistical Approaches to Language. MIT Press, Cambridge.

Dessus, P., B. Lemaire, et al. (2000). Free text assessment in a virtual campus. Proceedings of the 3rd International Conference on Human System Learning.

Dikli, S. (2006). "An overview of Automated Scoring of Essays." Journal of Technology, Learning and Assessment **5**(1).

Downey, D., M. Broadhead, et al. (2005). Locating Complex Named Entities in Web Text. Proceedings of IJCAI.

Ebel, R. L. (1979). Essentials of educational measurement. Englewood Cliffs, New Jersey, Prentice-Hall.

Elliot, S. (2003). IntelliMetric: from here to validity. . Automated essay scoring: A cross disciplinary approach. M. D. S. a. J. C. Burstein. Mahwah, NJ, Lawrence Erlbaum Associates.

Elliot, S. and C. Mikulas (2004). The impact of MY Access! use on student writing performance: A technology overview and four studies from across the nation. . Annual Meeting of the National Council on Measurement on Education, April 12-16. San Diego, CA.

Enokizu, H., S. Murakam, et al. (2008). Automatic Extraction of Important Sentences from Story Based on Connecting Patterns of Propositions. 7th WSEAS Int. Conference on Artificial Intelligence, Knowledge Engineering and Data Bases. University of Cambridge, UK.

Etzioni, O., M. Cafarella, et al. (2005). "Unsupervised Named-Entity Extraction from the Web: An Experimental Study." Artifical Intelligence **165**: 91-134.

Fleischman, M. and E. Hovy (2002). Fine Grained Classification of Named Entities. Conference on Computational Linguistics.

Florian, R., A. Ittycheriah, et al. (2003). Named Entity Recognition through classifier combination. Proceedings of CoNLL.

Foltz, P. W. (1996). "Latent Semantic Analysis for text-based research." Behavior Research Methods, Instruments and Computers **28**(2): 197-202.

Ford (2000). "Automated scoring of writing assessments." Powerpoint presentation for the Assessment Council.

Galliers, R. D. (1991). Choosing appropriate information systems research methodologies. North-Holland, Amsterdam, Nissen, HE, Klein, HK & Hirschheim, R.

Garcia, E. (2007) "SVC and LSI Tutorial 1: Understanding SVD and LSI." **Volume**, DOI:

Girju, R. and D. Moldovan (2002). Text Mining for Causal Relations. Proceedings of FLAIRS-02, American Association for Artificial Intelligence.

Godshalk, F., F. Swineford, et al. (1996). The measurement of writing ability. College Entrance Examination Board. New York.

Green, N. L. (2002). "Designing an Ontology for Artificial Intelligence in the Narrative Arts." AAAI Technical Report: 39-40.

Greimas, A. J. (1989). "The cognitive dimension of narrative discourse." New literary history **20**(3): 563.

Gruber, T. R. (1993). "A Translation Approach to Portable Ontology Specification." Knowledge Acquisition **5**: 199-220.

Guba, E. G. and Y. S. Lincoln (1989). Fourth Generation Evaluation, NewBury Park, CA: Sage.

Hamp-Lyons, L. (2001). Fourth generation writing assessment. Mahwah, New Jersey, Lawrence Erlbaum Associates.

Hearst, M. (2000). "The debate on automated essay grading." IEEE intelligent systems(15): 22-37.

Hevner, A. R., S. T. March, et al. (2003). "Design Science in Information Systems Research." MIS Quarterly.

Jahn, M. (2003). "Narratology: A Guide to the Theory of Narrative." English Department, University of Cologne.

Jones, E. R. R. (1999). Automatically Generating Extraction Patterns from Untagged Text. Proceedings of the Sixteenth National Conference on Articial Intelligence, The AAAI Press/MIT Press.

Juristo, N. and A. M. Moreno (2002). "Reliable knowledge for software development." Software IEEE **19**(5): 98-99.

Kukich, K. (2000). "Beyond automated essay scoring." IEEE intelligent systems **15**(5): 22.

Labov, W. (1972). Language in the Inner City.

Landauer, T. K., D. Laham, et al. (2000). "The Intelligent Essay Assessor." IEEE intelligent systems **15**(5): 27-31.

Landauer, T. K., D. Laham, et al. (2003). Automated Essay Scoring and annotation of essays with the Intelligent Essay Assessor. Automated Essay Scoring: a cross-disciplinary perspective. M. a. B. J. C. Shermis, Lawrence Erlbaum**:** 87-112.

Landauer, T. K., P. W. Foltz, et al. (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 21st annual international ACM SIGIR conference on Research and development in information retrieval. Melbourne, Australia**:** 90-95.

Leacock, C. and M. Chodorow (2004). "C-rater: Automated Scoring of Short-Answer Questions." COMPUTERS AND THE HUMANITIES **37**(4): 389-405.

Lucariello, J. (1990). Canonicality and consciousness in child narrative. Narrative thought and narrative language., Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc**:** 131-149.

Maedche, A. and S. Staab (2009). "Ontology learning for the semantic web." Intelligent Systems, IEEE **16**(2): 72-79.

Mason, O. and I. Grove-Stephenson (2002). Automated Free Text Marking with Paperless School. Sixth International Computer Assisted Assessment Conference, Loughborough University, Loughborough, UK.

Mikheev, A. (1999). A knowledge-free method for capitalized word disambiguation. Proceedings of the 37th annual meeting of the ACL, College Park, Maryland, Accosiation for Computational Linguistics.

Ming, P. Y., A. A. Mikhailov, et al. (2000). Intelligent essay marking system. Learners Together. NgeeANN Polytechnic, Singapore.

Mitchell, T., T. Russel, et al. (2002). Towards Robust Computerised Marking of Free-Text Responses. Sixth International Computer Assisted Assessment Conference, Loughborough University, Leicestershire UK.

Nadeau, D. and S. Sekine (2007). "A survey of Named Entity Recognition and Classification." Sekine, S. and Ranchod, E. Named Entities:

Recognition, classsification and use. Special issue of Lingvisticae Investigationes **30**(1): 3-26.

Nadeau, D., P. Turney, et al. (2006). Unsupervised Named Entity Recognition: Generating Gazetteers and Resolving Ambiguity. 19th Conference on Artificial Intelligence. Canada.

National Assessment Program (2010). - Literacy and Numeracy. D. o. Education. Western Australia, Australian Curriculum Assessment and Reporting Authority.

Nedjalkov, V. P. and G. Silnickij (1973). "The topology of causative constructions." Folia linguistica **6**(3-4): 273.

Nouioua, F. (2008). A Hueristic Approach to Order Events in Narrative Texts. 15th International Symposium on Temporal Representation and Reasoning.

Nunamaker, J. F., M. Chen, et al. (1991). "Systems Development in Information Systems Research." Journal of Management Information Systems **7**: 89-106.

Page, E. B. (1966). "The imminence of grading essays by computers." Phi Delta Kappan **47**: 238-243.

Page, E. B. and N. S. Petersen (1995). "The Computer Moves into Essay Grading: Updating the Ancient Test." Phi Delta Kappan **76**.

Palincsar, A., A. Brown, et al. (1994). Language learning disabilities in school-aged children and adolescents. Models and practices of dynamic assessment. Boston, Allyn and Bacon.

Palmer, D., D. and D. Day, S. (1997). A statistical profile of the Named Entity Task. Proceedings of the 5th conference on Applied Natural Language Processing, Washington, DC, Accosiation for Computational Linguistics.

Palmer, J., R. Williams, et al. (2003). On-line assessment and free-response input - a pedagogic and technical model for squaring the circle. Proceedings of the 7th Computer Assisted Assessment Conference.

Perez-Marin, D., I. Pascual-Nieto, et al. (2009). "Computer-assisted assessment of free-text answers." The Knowledge Engineering Review **24**(4): 353–374.

Petasis, G., F. Vichot, et al. (2001). Using machine learning to maintain rule-based named-entity recognition and classification systems. Proceedings of the 39th Annual Meeting on Association for Computational Linguistics Toulouse, France

Poibeau, T. and L. Kossien (2001). Proper Name Extraction from Non-Journalistic Text. Computational Linguistics in the Netherlands Meeting. W. Daelemans, K. Sima'an, J. Veenstra and J. Zavrel. New York**:** 144-157.

Powers, D. E., J. Burstein, et al. (2002). "Stumping e-rater: challenging the validity of automated essay scoring." Computers in Human Behavior **18**: 103-134.

Rajman, M. and R. Besancon (1997). Text Mining: Natural Language Techniques and Text Mining Applications. 7th IFIP 2.6 Working Conference on Database Semantics Leysin, Chapman and Hall.

Rittel, H. J. and M. M. Webber (1984). Planning Problems Are Wicked Problems. New York, John Wiley & Sons.

Rudner, L. M. (2001). An overview of three approaches to scoring written essays by computer.

Rudner, L. M. and T. Liang (2002). "Automated essay scoring using Bayes' Theorem." The Journal of Technology, Learning and Assessment **1**(2): 3-21.

Rudner, L. M., V. Garcia, et al. (2006). "An Evaluation of the IntelliMetric." Journal of Technology, Learning and Assessment **4**(4).

Scharfe, H. (2004). Narrative Ontologies. Proceedings of Knowledge Economy Meets Science.

Sekine, S. (1998). Nyu: Description of the Japanese NE system used for Met-2. Proceedings of Message Understanding Conference.

Sekine, S., K. Sudo, et al. (2002). Extended Named Entity Hierarchy. Proceedings of the LREC

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

Shermis, M. D. and F. D. Barrera (2002). "Exit assessments: Evaluating writing ability through Automated Essay Scoring " Education Resources Information Center ED464950.

Shneiderman, B. (1997). "Designing information-abundant web sites: Issues and recommendations."

Streeter, L., K. Pstoka, et al., Eds. (2003). The credible grading machine: Automated essay scoring in the dod.

Trabesso, T. and L. Sperry (1985). "Causal relatedness and importance of story events." Journal of memory and language **24**(5): 595.

Tuffield, M. M., D. E. Millard, et al. (2006). Ontological approaches to modelling narrative. Proceedings of 2nd AKT DTA Symposium.

Vaishnavi, V. and W. Kuechler (2004). "Design Research in Information Systems."

Valenti, S., F. Neri, et al. (2003). "An Overview of Current Research on Automated Essay Grading." Journal of Information Technology Education 2: 319-330.

Wang, J. and M. S. Brown (2007). "Automated Essay Scoring Versus Human Scoring: A Comparative Study." Journal of Technology, Learning and Assessment 6(2).

Warschauer, M. and P. Ware (2006). "Automated writing evaluation: defining the classroom research agenda." Language Teaching Research 10: 157-180.

Williams, R. (2001). Automated essay grading: an evaluation of four conceptual models. 10th Annual Teaching Learning Forum. Perth, Curtin University of Technology, Herrman A and Kulski M, Expanding Horizons in Teaching and Learning.

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

Witte, R., T. Kappler, et al. (2007). Ontology Design for Biomedical Text Mining. Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences, Springer US: 281-313.

Witten, I. H. (2004). Text mining, CRC Press.

Yang, Y., C. W. Buckendahl, et al. (2002). "A Review of Strategies for Validating Computer-Automated Scoring." Applied measurement in education 15(4): 391-412.

Plackett, R.L. (1983). "Karl Pearson and the Chi-Squared Test". International Statistical Review (International Statistical Institute (ISI)) 51 (1): 59–72

Skoog, D.A., et al., eds. Fundamentals of Analytical Chemistry. 8 ed. 2003, Brooks Cole: Belmont.

Appendix A - Physical and Mental State Checklist

Physical States	Mental States				
pain	adored	defeated	happy	nice	surprised
numb	afraid	dejected	hassled	numb	suspicious
aching	aggravated	delighted	hateful	optimistic	sympathetic
broken	agitated	depress	helpless	outraged	tense
ill				overwhelmed	
safe	agonized	desired	hesitant		terrified
unsafe		disappointed			
hurt	agony		homesick	panicky	thrilled
exhausted	alarmed	disgusted	hopeful	passionate	tired
	alienated	disliked	hopeless	pessimistic	tormented
	amazed	dismayed	horror	petrified	triumphant
	amused	distressed	horrible	pleased	troubled
					uncomfortable
	anger	disturbed	hostile	proud	
	angry	dreadful	humiliated	puzzled	uneasy
	anguish	eager	hysterical	queasy	unhappy
	annoyed	ecstatic	impatient	rageful	unsettled
	antsy	edgy	indifferent	raptured	upset
	anxious	elated	infatuated	regretful	vengeful
	apprehensive	embarrassed	inferior	rejected	vicious
	aroused	enraged	insecure	relieved	weary
	ashamed	enthralled	insulted	reluctant	woeful
	astonished	enthused	irate	remorseful	worried
	attracted	envious	irked	resentful	worry
	awful	euphoric	irritated	restless	wrathful
		exasperated			
	awkward		isolated	revulsed	zealless
	bashful	excited	jealous	ridiculous	zestless
	bewildered	fatigued	jittery	riled	
	bitter	fear	jolly	rushed	
	blissed	fearful	joy	sad	
	bored	ferocious	joyous	satisfied	
	calm	fidgety	leery	scared	

	cautious	frantic	liked	scornful
	cheerful	frightened	loathe	secure
	concerned	frustrated	lonely	sensitive
	confident	furious	loving	shaky
	confused	glad	mad	shock
			melancholica	
	contempt	gleeful	l	shocked
	content	gloomy	miserable	shy
		griefstricke		
	critical	n	moody	silly
	curious	grouchy	mortified	sleepy
	cynical	grumpy	neglected	spiteful
	daydream	guilty	nervous	stressed

Appendix B -Scoring Logic Source Code

Audience

```
package src.p;


public class Audience

{

	public static void main(String [] args)

	{

		TXTFile f = new

TXTFile("C:\\Sean's_Work_Stuff_02\\WorkSpace\\AEG\\src\\audience.txt");

		for(int i=0;i<90;i++)

				{

					int noOfEvents = (f.parseEvents()[i]);

					int essayLength =

(f.parseNumSentence()[i]);

					boolean PMState = (f.parsepmState()[i]);

					double ratio = (f.parseRatio()[i]);

					String essayName = (f.parseName()[i]);

					int adj = (f.parse2()[i][1]);

					String audience = null;

					String grp = null;

					int noun = f.parse2()[i][2];

					double score = 5;


					//check Essay Length
```

```
                                    if (essayLength >= 30)
                                          score = score + 1;
                                    else if (essayLength > 24)
                                          score = score + 0.5;
                                    else if (essayLength > 14)
                                          score = score - 0.5;
                                    else if (essayLength > 9)
                                          score = score - 1;
                                    else
                                          score = 0;
                                    //check Events and Ratio
                                    if (noOfEvents >1)
                                    {
                                          if (ratio >= 0.30 && ratio <= 0.39
|| ratio >= 0.60 && ratio <= 0.85 || ratio > 0.50 && ratio < 0.59)
                                                {
                                                if (noOfEvents > 15)
                                                      score = score + 1.5;
                                                else if (noOfEvents >13 )
                                                      score = score + 1;
                                                else if (noOfEvents >= 10)

                                                      score = score - 0.5;

                                                else if (noOfEvents >= 8)
                                                      score = score - 1;

                                                else if (noOfEvents >= 5)
```

```
                        score = score - 3.5;
                    else
                        score = score - 5;
                    }
                }
            else
                {
                    //double check Events
                if (noOfEvents > 15)
                    score = score + 1;
                else if (noOfEvents >13 )

                    score = score + 0.5;

                else if (noOfEvents >= 10)

                    score = score - 1;

                else if (noOfEvents >= 8)

                    score = score - 1.5;

                else if (noOfEvents >= 5)

                    score = score - 4;
                else
                    score = score - 6;
```

```
                                    }
                                //check nouns
                                if (noun >50)
                                      score = score + 1;

                                //check grouping
                                if (essayLength < 9)

                                      grp= "c";

                                else
                                {
                                      if (ratio > 0.35 && ratio < 0.85){
                                if (essayLength > 26 && PMState == true)
                                            grp = "a";
                                      else
                                            grp = "b";
                                            }
                                      else
                                                  grp = "b";

                                }
                                //calculate score
                                if (grp == "a" && PMState == true||score
>= 8.5||score >=7 && grp == "b" && PMState == true)
                                      audience = "Good";

                                else if (score >= 5 && grp == "a" ||
```

```
score >= 6 && grp == "b"||score >= 7 && grp == "c")
                                        audience = "Intermediate";


                                  else
                                        audience = "Poor";


                                  System.out.println(essayName + "\t" + grp
+ "\t" +audience + "\t" + score + "\t " + essayLength + "\t "
                                              + noOfEvents + "\t" + ratio +
"\t" + adj+ "\t" + PMState+ "\t" + noun );
                        }
      }
}
```

Ideas

```
package src.p;


public class Ideas {


      public static void main (String [] args){
            TXTFile f = new
TXTFile("C:\\Sean's_Work_Stuff_02\\WorkSpace\\AEG\\src\\audience.txt");
            for(int i=0;i<90;i++)
            {
                  int noOfEvents = f.parseEvents()[i];
                  int essayLength = f.parseNumSentence()[i];
```

```
                    double ratio = f.parseRatio()[i];
                    String essayName = f.parseName()[i];
                    int adj = f.parse2()[i][1];
                    int adv = f.parse2()[i][3];
                    double score = 10;
                    String ideas = null;
                    String grp = null;

                    //check number of Events and Ratio
                    if (noOfEvents >1)
                    {
                                if (ratio >= 0.30 && ratio <= 0.39 ||
ratio >= 0.60 && ratio <= 0.85 || ratio > 0.50 && ratio < 0.59)
                                {
                                      if (noOfEvents > 15)
                                            score = score + 0;
                                      else if (noOfEvents >13 )

                                            score = score - 1.5;

                                      else if (noOfEvents >= 10)

                                            score = score - 2.5;

                                      else if (noOfEvents >= 8)

                                            score = score - 3;
```

```
                else if (noOfEvents >= 5)

                        score = score - 3.5;
                else
                        score = score - 5;
        }
        else
        {
                //double check Events
                if (noOfEvents >= 15)

                        score = score - 1;

                else if (noOfEvents >=12 )

                        score = score - 2;

                else if (noOfEvents >= 10)

                        score = score - 2.5;

                else if (noOfEvents >= 8)

                        score = score - 3.5;

                else if (noOfEvents >= 5)

                        score = score - 5.5;
```

```
                        else
                              score = score - 6;
                  }

            }
            else score = score - 10;

            //check Essay Length
            if(essayLength > 1){
                  if (essayLength > 30)
                        score = score + 0;

                  else if (essayLength >= 25)
                        score = score - 2;
                  else if (essayLength >15)
                        score = score - 2.5;
                  else if (essayLength > 9)
                        score = score -3.5;
                  else
                        score = score - 5;
            }
            else
                  score = score - 8;

            //check number of adjectives
            if(adj > 1){
                  if(adj >=20)
                        score =score + 0;
```

```
                else if (adj > 15)
                        score = score - 1;

                else if (adj > 10)
                        score = score - 1.5;

        }
        else
                score = score - 2;

        //check number of adverbs
        if(adv > 1){
                if(adv >=20)
                        score =score + 0;
                else if (adv > 15)
                        score = score - 1;

                else if (adv > 10)
                        score = score - 1.5;

        }
        else
                score = score - 2;

        //second checking stage
        if (essayLength > 30)
        {
                if (ratio >= 0.30 && ratio <= 0.39 || ratio >=
```

```
0.65 && ratio <= 0.85 || ratio > 0.50 && ratio < 0.59){
                              if(adj > 20 && adv > 20)
                                    grp = "a";
                              else
                                    grp = "b";
                        }
                        else
                              grp = "b";
                  }
                  else if (essayLength <= 30 && essayLength > 8)
                        grp = "b";
                  else if (essayLength <= 8)
                        grp = "c";




                  //calculate score
                  if (score >=7 && grp == "a" || score >= 8.5)
                        ideas = "Good";
                  else if (score <= 8 && score > 4 && grp == "a" ||
score <= 8 && score > 4 && grp == "b")
                        ideas = "Intermediate";
                  else if (score <= 4 && grp == "b" || score <= 4 &&
grp == "c")
                        ideas = "Poor";
                  System.out.println(essayName+ "\t" + grp + "\t" +
ideas+ "\t" + score + "\t" + essayLength  + "\t" + noOfEvents + "\t" +
ratio+ "\t" + adj+ "\t" + adv)
```

Character and Setting

```
package src.p;


public class CharacterandSetting {


	public static void main (String [] args)
	{
		TXTFile f = new
TXTFile("C:\\Sean's_Work_Stuff_02\\WorkSpace\\AEG\\src\\audience.txt");
		for(int i=0;i<90;i++)
		{
			int noOfEvents = (f.parseEvents()[i]);
			int essayLength = (f.parseNumSentence()[i]);
			double ratio = (f.parseRatio()[i]);
			String essayName = (f.parseName()[i]);
			int simpleCon = (f.parseSimple()[i]);
			int advCon = (f.parseAdvance()[i]);
			int adj = (f.parse2()[i][1]);
			int adv = f.parse2()[i][3];
			String cs = null;
			String grp = null;
			boolean PMState = (f.parsepmState()[i]);
			double score = 3;


			//Check number of simple connectives (4)
			if (simpleCon != 0)
			{
				if (simpleCon >=10){
```

```
                        score = score + 3;
                    }
                    else if (simpleCon >= 5)
                    {
                        score = score +1.5;
                    }
                    else if (simpleCon >= 1)
                    {
                        score = score + 0.5;
                    }
                }
                else
                    score = score + 0;

                //Check Number of Advanced Connectives (2)
                if (advCon >=1){
                    score = score + 2;
                }

                //Check Ratio and Events
                if (noOfEvents >1)
                {
                        if (ratio >= 0.30 && ratio <= 0.39 ||
ratio >= 0.60 && ratio <= 0.85 || ratio > 0.50 && ratio < 0.59)
                        {
                            if (noOfEvents > 15)
                                score = score + 0;
                            else if (noOfEvents >13 )
```

```
                                    score = score - 0.5;

                              else if (noOfEvents >= 10)

                                    score = score - 1;

                              else if (noOfEvents >= 8)

                                    score = score - 1.5;

                              else if (noOfEvents >= 5)

                                    score = score - 3.5;
                              else
                                    score = score - 5;
                        }
                        else
                        {
                              //double check Events
                              if (noOfEvents >= 1)
                                    {     if (noOfEvents > 15)
                                                score = score +
1;
                                          else if (noOfEvents >
10)
                                                score = score +
0.5;
```

```
                                        else if (noOfEvents >
5)
                                              score = score -
4;
                                        else
                                              score = score -5;
                                  }
                                  else
                                        score = score - 8;
                            }
                      }
                      else score = score - 10;



                      //check Essay Length
                      if (essayLength >= 30)
                            score = score + 1;
                      else if (essayLength > 24)
                            score = score + 0.5;
                      else if (essayLength > 14)
                            score = score - 0.5;
                      else if (essayLength > 9)
                            score = score - 1;
                      else
                            score = 0;


                      //check number of adjectives
                      if (adj >=1)
```

```
                {if (adj > 20)
                        score = score + 1;
                else if (adj > 15)
                        score = score + 0.5;
                else if (adj >= 10)
                        score = score - 1;
                else if (adj > 5)
                        score = score - 4;
                }
        else
                score = score - 8;
        //check number of adverbs
        if(adv > 1){
                if(adv >=20)
                        score =score + 1;
                else if (adv >= 15)
                        score = score + 0.5;

                else if (adv >= 10)
                        score = score + 0;
                else if (adv >= 6)
                        score = score - 2;
                else
                        score = score - 6;
        }
        else
                score = score - 8;
```

```
                    //cap score

                    if (score > 10)
                          score = 10;

                    //check grouping
                    if(essayLength >= 5 && simpleCon >=5)
                    {
                                if(essayLength >= 28)
                                {
                                      if (ratio > 0.35 && ratio < 0.85)
                                      {
                                            if(simpleCon > 9 && advCon >=
1 && PMState == true)
                                            grp = "a";
                                            else
                                            grp = "b";
                                      }
                                      else
                                            grp = "b";
                                }
                                else
                                      if (essayLength < 28 && simpleCon
>=7 && advCon >=1)
                                            grp = "b";
                                      else
                                            grp = "c";
                    }
```

```
                    else
                          grp = "c";
                    //Calculate Score
                    if (score >= 8.5 && grp =="a"||score > 8.5 && grp ==
"b" && advCon >=1 && PMState == true)
                          cs = "Good";
                    else if (score <= 4 && PMState == false || score <= 5
&& grp == "c")
                          cs = "Poor";
                    else
                          cs = "Intermediate";


                    System.out.println(essayName + "\t" + grp + "\t" +cs
+ "\t" + score + "\t " + essayLength + "\t " + noOfEvents + "\t" +
ratio + "\t" + adj
                                      + "\t " + adv+ "\t " + PMState);
              }
      }
}
```

Cohesion

```
package src.p;


public class Cohesion {
      public static void main (String [] args){
            TXTFile f = new
TXTFile("C:\\Sean's_Work_Stuff_02\\WorkSpace\\AEG\\src\\audience.txt");
```

```
for(int i=0;i<90;i++)
{
	int noOfEvents = (f.parseEvents()[i]);
	int essayLength = (f.parseNumSentence()[i]);
	double ratio = (f.parseRatio()[i]);
	String essayName = (f.parseName()[i]);
	int simpleCon = (f.parseSimple()[i]);
	int advCon = (f.parseAdvance()[i]);
	int adj = (f.parse2()[i][1]);
	int adv = f.parse2()[i][3];
	String cohesion = null;
	String grp = null;
	double score = 3;

	//Check number of simple connectives (4)
	if (simpleCon != 0)
	{
		if (simpleCon >=10)
			score = score + 3;
		else if (simpleCon >= 5)
			score = score +1.5;
		else if (simpleCon >= 1)
			score = score + 0.5;
	}
	else
		score = score + 0;

	//Check Number of Advanced Connectives (2)
```

```
                    if (advCon >=1)
                          score = score + 2;


                    //Check Ratio and Events
                    if (noOfEvents >1)
                    {
                                if (ratio >= 0.30 && ratio <= 0.39 ||
ratio >= 0.60 && ratio <= 0.85 || ratio > 0.50 && ratio < 0.59)
                                {
                                      if (noOfEvents > 15)
                                            score = score + 0;
                                      else if (noOfEvents >13 )

                                            score = score - 0.5;

                                      else if (noOfEvents >= 10)

                                            score = score - 1;

                                      else if (noOfEvents >= 8)

                                            score = score - 1.5;

                                      else if (noOfEvents >= 5)

                                            score = score - 3.5;
                                      else
                                            score = score - 5;
```

```
                        }
                        else
                        {
                              //double check Events
                              if (noOfEvents >= 1)
                                    {     if (noOfEvents > 15)
                                                score = score +
1;
                                          else if (noOfEvents >
10)
                                                score = score +
0.5;
                                          else if (noOfEvents >
5)
                                                score = score -
4;
                                          else
                                                score = score -5;
                                    }
                                    else
                                          score = score - 8;
                        }
                  }
                  else score = score - 10;

                  //check Essay Length
                  if (essayLength >= 30)
                        score = score + 1;
```

```
else if (essayLength > 24)
      score = score + 0.5;
else if (essayLength > 14)
      score = score - 0.5;
else if (essayLength > 9)
      score = score - 1;
else
      score = 0;

//check number of adjectives
if (adj >=1)
      {if (adj > 20)
            score = score + 1;
      else if (adj > 15)
            score = score + 0.5;
      else if (adj >= 10)
            score = score - 1;
      else if (adj > 5)
            score = score - 4;
      }
else
      score = score - 8;
//check number of adverbs
if(adv > 1){
      if(adv >=20)
            score =score + 1;
      else if (adv >= 15)
            score = score + 0.5;
```

```
                        else if (adv >= 10)
                                score = score + 0;
                        else if (adv >= 6)
                                score = score - 2;
                        else
                                score = score - 6;
                }
                else
                        score = score - 8;

                //cap score

                if (score > 10)
                        score = 10;

                //check grouping
                if(essayLength >= 5 && simpleCon >=5)
                {
                        if(essayLength >= 28)
                        {
                                if (ratio > 0.35 && ratio < 0.85)
                                {
                                        if(simpleCon > 9 && advCon >=
1)

                                        grp = "a";
                                        else
                                        grp = "b";
```

```
                                    }
                                    else
                                          grp = "b";
                              }
                              else
                                    if (essayLength < 28 && simpleCon
>=7 && advCon >=1)
                                          grp = "b";
                                    else
                                          grp = "c";
                  }
                  else
                        grp = "c";

                  //Calculate Score
                  if (score >= 8.5 && grp =="a"||score > 8 && grp ==
"b" && advCon >=1 || score == 10)
                        cohesion = "Good";
                  else if (score < 3.5 || score < 2.5 && grp == "c")
                        cohesion = "Poor";
                  else
                        cohesion = "Intermediate";

                  System.out.println(essayName+ "\t" + grp + "\t" +
cohesion+ "\t" + score + "\t" + simpleCon + "\t" + advCon
                              + "\t " + essayLength + "\t" + ratio+
"\t" + adj);
```

Appendix C - List of Connectives

Simple	Advanced	
already	above all	secondly
also	additionally	sequencing
among	afterwards	significantly
and	alternatively	similarly
as with	although	stemming from this
because	an upshot of	straightaway
before	apart from	therefore
below	as a result	these include
clearly	as exemplified by	throughout
during	as long as	whenever
earlier	as revealed by	whereas
except	as well as	
first	beneath	
firstly	beyond	
from	by the time	
hence	comparing	
if	consequently	
indeed	contrasting	
inside	despite	

into	emphasising		
later	equally		
like	especially		
moreover	finally		
near	for example		
next	for instance		
now	furthermore		
on	hitherto		
out of	illustrating		
outside	in addition		
second	in other words		
since	in particular		
so	in that respect		
such as	in the case of		
then	including		
thus	instead of		
till	lastly		
to	likewise		
too	meanwhile		
towards	notably		
unless	on the contrary		
unlike	on the other hand		

until within yet	otherwise placing qualifying respects	

Appendix B - Event Detection Source Code

```
package org.debii.aeg.sean.nlp;



import java.io.File;

import java.io.IOException;

import java.util.ArrayList;

import java.util.HashMap;

import java.util.List;

import java.util.Map;

import java.util.Properties;

import java.util.regex.Matcher;

import java.util.regex.Pattern;

import java.text.DecimalFormat;



import org.apache.log4j.Logger;

import org.debii.aeg.anomalousFilter.ReadDocs;

import org.debii.aeg.entity.Pos;

import org.debii.aeg.entity.State;

import org.debii.aeg.utilities.GetPath;

import org.debii.aeg.utilities.ReadMentalStateList;
```

```
import org.debii.aeg.utilities.Utilities;


import edu.stanford.nlp.ling.CoreAnnotations;
import edu.stanford.nlp.ling.CoreLabel;
import edu.stanford.nlp.ling.CoreAnnotations.PartOfSpeechAnnotation;
import edu.stanford.nlp.ling.CoreAnnotations.SentencesAnnotation;
import edu.stanford.nlp.ling.CoreAnnotations.TextAnnotation;
import edu.stanford.nlp.ling.CoreAnnotations.TokensAnnotation;
import edu.stanford.nlp.pipeline.Annotation;
import edu.stanford.nlp.pipeline.DefaultPaths;
import edu.stanford.nlp.pipeline.StanfordCoreNLP;
import edu.stanford.nlp.util.CoreMap;


public class AnnotateEssays {
	private static Logger
logger=Logger.getLogger("org.debii.aeg.sean.nlp.AnnotateEssays");
	private String fileName;
	/*the physic and mental state List file URL*/
	private String url=GetPath.getMentalStateFilePath();
	private boolean hasMentalPhysicState=false;
	private boolean hasActorState=false;
	private boolean hasActionState=false;
```

```
private boolean hasLocationState=false;

private boolean isNameAppearFirstTime=false;

private ArrayList<String> personList=new ArrayList<String>();

private String debugInfo="";


public AnnotateEssays(){

        resetStateIndicator();

}


public void setFileName(String f){

        fileName=f;

}


public Map<String, String> getNerOverwrite() {

        return nerOverwrite;

}


public void setNerOverwrite(Map<String, String> nerOverwrite) {

        this.nerOverwrite = nerOverwrite;

}


//NE to be rewritten upon pattern.
```

```
private Map<String,String> nerOverwrite;



public Annotation myAnnotate(String inputStr){
        String conllURL=GetPath.getNERClassifierFilePath();
        Properties props = new Properties();
        props.put("annotators", "tokenize,ssplit,pos,lemma,ner");
        props.put("ner.model.MISCclass", conllURL);

        DefaultPaths.DEFAULT_NER_CONLL_MODEL="afebde";
        StanfordCoreNLP pipeline= new StanfordCoreNLP(props);
        Annotation document =new Annotation(inputStr);
        pipeline.annotate(document);



        return document;
}



/*This method will testing given set of text and going to evaluate the pos
 * for each words. By doing this will help user to detect possible POS set
 * in order to improve the accuracy of pos detection
```

```
	 * @param inputStr testing string
	 * */
	public void viewPatterns(String inputStr){
		Annotation document=myAnnotate(inputStr);
		String excelDelimiter=Pos.connector;
		String msg="";
		List<CoreMap> sentences = document.get(SentencesAnnotation.class);
		for(CoreMap sentence:sentences){
			String A="";
			String B="";
			for(CoreLabel token: sentence.get(TokensAnnotation.class)){
				String word=token.get(TextAnnotation.class);
				if(word.equals("."))continue;
				String pos=token.get(PartOfSpeechAnnotation.class);
				String
ne=token.get(CoreAnnotations.NamedEntityTagAnnotation.class);
				A+=word+excelDelimiter;
				B+=pos+"|"+ne+excelDelimiter;
			}
				A+="\n";
				B+="\n";
```

```
                msg+=A+B+"\n";
            }
            logger.info(msg);
        }

        /*This method will annotate the given essay with POS and NE, and then detecting if there is
         * a event in each sentence according to the event matching pattern given as parameter.
         *
         * @param inputStr which is the essays content in string format
         * @param event detection pattern which is combination of regex
         * @param neOverWrite: Map<String,String> which is regex pattern and UserDefined NE to overWrite Stanford NE
         * @return annotateEssay: event detection report
         * */
        public String annotateEssays(String inputStr,String patternStr,
                Map<String,String> neOverWrite,boolean writeTofile){
            String annotatedEssay="\nWord\t|\tPOS\t|\tNER\t\n--------------------------------------------------------------------------------\n";
            String excelDelimiter="#";
            this.nerOverwrite=neOverWrite;
```

```
                String counterDesc="";
                int counterEvent=0;
                int notEventCounter=0;
                Annotation document=myAnnotate(inputStr);
                List<CoreMap> sentences = document.get(SentencesAnnotation.class);
                String
debugReport="SentenceNo"+excelDelimiter+"Actor"+excelDelimiter+"Action"+excelDel
imiter+"Location|isEvent"+excelDelimiter+"\n";
                int sentenceCount=0;
                for(CoreMap sentence:sentences){
                        resetStateIndicator();
                        sentenceCount++;

        annotatedEssay+=sentenceCount+"\t"+sentence.toString()+"\n==========
==========+\n";
                        //logger.warn(annotatedEssay);
                        String annotedSentence="";
                        boolean isEvent=true;
                        for(CoreLabel token: sentence.get(TokensAnnotation.class)){
                                String word=token.get(TextAnnotation.class);
                                String pos=token.get(PartOfSpeechAnnotation.class);
                                String
```

```
ne=token.get(CoreAnnotations.NamedEntityTagAnnotation.class);
                                String
text=word+Pos.wordDelimiter+pos+"|"+ne+Pos.connector;
                                annotedSentence+=text;
                        }
                        debugReport+=sentenceCount+excelDelimiter;
                        String[] results=this.parsePatterns(patternStr, annotedSentence);
                        //example:a@DT|O-beautiful@JJ|O-summers@NNS|O
%c.location-## XXXXX

                        for(int i=0;i<results.length;i++){
                                String result=results[i];
                                System.out.println("["+i+"] "+result);
                                logger.warn("if result is empty then is not event~
"+result);
                                if(!(result.trim()).equalsIgnoreCase("")){
                                        String[] resultTokens=result.split(Pos.tokenSplit);
                                        for(String resultToken:resultTokens){
                                                String word=resultToken;
                                                debugReport+=word;
                                        }
                                        /*temporary to see which state exactly is true
```

```
                                        * will be changed if more states added later. */
                                        if(i==0){
                                                this.hasActorState=true;
                                        }else if(i==1){
                                                this.hasActionState=true;
                                        }else if(i==2){
                                                this.hasLocationState=true;
                                        }
                                        logger.warn(this.hasActorState+"                and
"+this.hasActionState+" and  "+this.hasLocationState);
                                }else{
                                        isEvent=false;
                                }
                                debugReport+=excelDelimiter;
                        }


                        String token=results[results.length-1];
                        System.out.println("Token                                   is
"+token+"====================size is "+results.length);
                        String[] tokens=token.split(Pos.connector);
                        boolean mentalStateDetected=false;
                        String tempType=null;
```

```
                    for(int i=0;i<tokens.length;i++){
                            /*adding the mental or physic state conditions
                             * added on 17/05/2011
                             * */
                            String temp=tokens[i];
                            logger.debug("temp is ============="+temp);
                            annotatedEssay+=printTuples(temp);
                    }

                    /*here will check is the isEvent is false, and the reason cause it to
                     * be false is because the hasLocationState is not found, but the
                     * hasMentalPhysicState is true, than the overall event will
                     * be set back to true*/
                    logger.warn("isEvent is"+isEvent);
                    if(isEvent==false){
                            logger.warn("hasLocationState                    is
"+this.hasLocationState);
                            logger.warn("hasMentalPhysicState                is
"+this.hasMentalPhysicState);
                            if(this.hasLocationState==false                  &&
this.hasMentalPhysicState){
                                    isEvent=true;
```

```
                                }
                                if(this.isNameAppearFirstTime){
                                        isEvent=true;

                                }
                                this.isNameAppearFirstTime=false;
                        }

                        if(isEvent){
                                counterEvent++;
                                counterDesc+=1+" ";
                                debugReport+="YES"+excelDelimiter;
                                annotatedEssay+="====================\n     isEvent:
YES\n"+"=========================\n";
                        }else{
                                notEventCounter++;
                                counterDesc+=0+" ";
                                debugReport+="NO"+excelDelimiter;
                                annotatedEssay+="====================\n     isEvent:
NO\n"+"=========================\n";
                        }
                        //logger.info("wpns size is "+wpns.size());
```

```
                    debugReport+="\n";
                }
                logger.info(debugReport);
                System.out.println(debugReport);
                logger.info(annotatedEssay);
                int totalEvent=counterEvent+notEventCounter;
                double percent  = (counterEvent/(double)totalEvent)*100;
                DecimalFormat formatter = new DecimalFormat("0.00");
                String headerInfo="\n EventCounter: "+counterEvent+" and
NotEventCounter: "+notEventCounter+"\n"+
                "Event Sequence: "+counterDesc+"\nEvent
Ratio"+counterEvent+"/"+totalEvent+"="+formatter.format(percent)+"%";
                annotatedEssay=headerInfo+annotatedEssay;
                if(writeTofile){
                    this.printToFile(fileName, fileName+"_import.txt",debugReport);
                    this.printToFile(fileName,
fileName+"_detailed.txt",annotatedEssay);
                }
                return annotatedEssay;
        }
```

```
        /*This is a helper method that will check if the directory is exist, if not create
this directory
         * @param folderURL, a directory URL
         * */
        private void checkDir(String folderURL){
                if(!new File(folderURL).exists()){
                        if(new File(folderURL).mkdirs()){
                                logger.debug("Message: Created Directory: "+folderURL );
                        }else{
                                logger.debug("Error: Created Directory: "+folderURL+"
Failed" );
                        }
                }
        }


        /*This method will write result into the given URL
         * @param folderName: which is the name of the essayFile Name
         * @param fileName: which is the filename, usually is FolderName_import.txt
for excel dataImport or
         * FolderName_details.txt for the essay annotate details
         * @param text: which is the actual text write to the file.
         * */
```

```
	public void printToFile(String folderName,String fileName, String text){
		String
base=GetPath.getEssayTestOutputRootPath()+folderName+File.separator;
		checkDir(base);
		Utilities.logPrint(base+fileName,text);
	}



	/*state indicator will be reset for next sentence
	 * processing
	 * */
	public void resetStateIndicator(){
		 hasMentalPhysicState=false;
		 hasActorState=false;
		 hasActionState=false;
		 hasLocationState=false;
	}

	private String composeReWriteNE(String token,String new_ne){
		String out="";
		String[] tokens=token.split(Pos.connector);
		for(int i=0;i<tokens.length-1;i++){
```

```
                        String t=tokens[i];
                        String[] innerTokens=t.split(Pos.delimiter);
                        out+=innerTokens[0]+"|"+new_ne+Pos.connector;
                }
                String t=tokens[tokens.length-1];
                String[] innerTokens=t.split(Pos.delimiter);
                out+=innerTokens[0]+"|"+new_ne;
                return out;
        }

/*example:a@DT|O-beautiful@JJ|O-summers@NNS|O-## XXXXX*/
        private String getWordFromToken(String token){
                String word=null;
                int i=token.indexOf("@");
                word=token.substring(0, i).toLowerCase().trim();

                return word;
        }
         /*example:a@DT|O-beautiful@JJ|O-summers@NNS|O-## XXXXX*/
        private String resetPyshicOrMentalNE(String token,String type){
                String word=null;
                logger.debug("token is resetPhysical state is "+token);
```

```
                int i=token.indexOf("|");
                logger.debug("substring0-i in resetPhysical state is "+token.substring(0,
i)+" and Type is "+type);
                if(i!=-1){
                        word=token.substring(0, i)+"|"+type;
                }
                return word;
        }



        /*this method will help to detect if the annotated token
         *match to the Physical State or Mental State list provided
         *if matched, then rewrite NE part of the annotated token with
         *MentalState or Physical State accordingly. if not found return null
         *@param token annotated word from essay
         *@return the updated version of annotated word with Physical or MentalState
         * */
        private String checkMentalPhysicState(String token){
                String newToken=null;


                ReadMentalStateList read=new ReadMentalStateList();
                try {
```

```
                read.processMentalPhysicList(url);
                ArrayList<String> physicList=read.getPhysicList();
                ArrayList<String> mentalList=read.getMentalList();

                for(String pList:physicList){
                        String word=this.getWordFromToken(token);
                        if(word.equals(pList)){
                                newToken=this.resetPyshicOrMentalNE(token,
"Physical State");
                        }
                }

                for(String mList:mentalList){
                        String word=this.getWordFromToken(token);
                        if(word.equals(mList)){
                                newToken=this.resetPyshicOrMentalNE(token,
"Mental State");
                        }
                }
        } catch (IOException e) {
                e.printStackTrace();
        }
```

```
                return newToken;
        }




        /*
         *
         * @param patternStr groups of regex Patterns connected with conncetors
         * @param token is the pos+ne series that need to be matched with the regex patterns.
         * @return Matched strings if there is no match then a empty string will be stored..
         * example:a@DT|O-beautiful@JJ|O-summers@NNS|O-## XXXXX
         * returned String array index 0 is actor state, 1 is action state, 2 is location state
         * 3 is copy of original token
         * */
        public String[] parsePatterns(String patternStr,String token){


                //get each pattern Actor:Action:Location
                @SuppressWarnings("unused")
                String state="";
```

```
                logger.warn(token);
                int verbCounter=this.getNumOfVerbs(token);
                String[] patterns=patternStr.split(Pos.tokenSplit);
                if(patterns==null){
                        throw new IllegalArgumentException();
                }
                //prepare the array that store the matched string
                //if there is no match for the given pattern, then an
                //empty string will be sotred. the last one is the altered senetence
                String[] matched=new String[patterns.length+1];
                for(int i=0;i<patterns.length;i++){
                        String temp=patterns[i];
                        String[] innerTokens=temp.split(Pos.or);
                        if(i==0){
                                logger.debug("Actor patterns is"+temp+" and pattern size
is "+innerTokens.length);
                        }else if(i==1){
                                logger.debug("Action patterns is"+temp+" and pattern
size is "+innerTokens.length);
                        }else if(i==3){
                                logger.debug("State patterns is"+temp+" and pattern size
is "+innerTokens.length);
```

```
                    }else if(i>4){
                            throw new IllegalArgumentException();
                    }
                    String result="";
                    for(int j=0;j<innerTokens.length;j++){
                            String innerToken=innerTokens[j];
                            if(i==0){
                                    logger.debug("Actor Inner Pattern is :
"+innerToken);
                            }else if(i==1){
                                    logger.debug("Action Inner Pattern is :
"+innerToken);
                            }else if(i==2){
                                    logger.debug("State Inner Pattern is :
"+innerToken);
                            }
                            Pattern pattern=Pattern.compile(innerToken);
                            logger.debug("Match agaist the token "+token);
                            Matcher matcher=pattern.matcher(token);
                            while(matcher.find()){
                                    debugInfo=matcher.group();
                                    logger.debug("matched word : "+debugInfo);
```

```
                                        String
matchedPattern=matcher.pattern().pattern();
                                        String actorName=this.getMyWord(debugInfo);
                                        checkPersonApprearFirstTime(matchedPattern,
actorName);
                                        if(debugInfo.equalsIgnoreCase("it@PRP|O")){
                                                continue;
                                        }
                                        /*this will test if the combine pattern of state
have verb in it.*/
                                        if(i==patterns.length-1){
                                                logger.warn("innerToken_State:
"+innerToken);

        if(this.getNumOfVerbs(debugInfo)>0&&verbCounter==1){

        //result+=debugInfo+Pos.tokenSplit;
                                                        logger.debug("pattern has verb
"+this.getNumOfVerbs(debugInfo));
                                                        continue;
                                                }
```

```
                                        String
tempNE=this.nerOverwrite.get(innerToken);
                                        logger.debug("tempNE is "+tempNE);
                                        if(tempNE!=null){
                                                String
newToken=this.composeReWriteNE(debugInfo, tempNE);
                                                logger.debug("before the matched
word is "+debugInfo);
                                                logger.debug("after rewriting the
matched word is "+newToken);

        token=token.replace(debugInfo,newToken);
                                                logger.debug("accumlated token is
"+token);
                                                debugInfo=newToken;
                                                result+=newToken+Pos.tokenSplit;
                                        }
```

```
                                        String[]
matchedStrs=this.debugInfo.split(Pos.connector);
                                        String tempType=null;
                                        boolean isMental=false;
                                        for(String str:matchedStrs){

                                                String
checkMental=this.checkMentalPhysicState(str);
                                                logger.debug("checkMental       for
"+str+" and is Mental is "+checkMental);
                                                if(checkMental!=null){
                                                        isMental=true;

        tempType=checkMental.substring(checkMental.indexOf("|")+1);
                                                        break;
                                                }
                                        }

                                        if(isMental){
                                                for(String str:matchedStrs){
```

```
            str=this.resetPyshicOrMentalNE(str, tempType);
                                                        String
newToken=this.composeReWriteNE(debugInfo, tempType);
                                                        logger.debug("before
rewriteing mentalstate is "+debugInfo);
                                                        logger.debug("after
rewriting mentalstate is "+newToken);


            token=token.replace(debugInfo,newToken);
                                                        logger.debug("after   mental
new accumulated token is  "+token);


            result+=newToken+Pos.tokenSplit;
                                                    }
                                                }



                                                logger.debug("regex                matched:
"+debugInfo);
                                                logger.debug("=-=regex result: "+result);
                                            }else{
                                                result+=debugInfo+Pos.tokenSplit;
```

```
                                        }

                                }

                        }

                        logger.warn("match["+i+"] is '"+result+"'");

                        matched[i]=result;

                        result="";

                }



                //now assign the whole sentence.

                matched[patterns.length]=token;

                System.out.println("###################mathced[length] is "+token);

//              String[] tokens=token.split(Pos.connector);

//              for(int i=0;i<tokens.length;i++){

//                      printTuples(tokens[i]);

//              }

//

                return matched;

        }
```

```
	public void checkPersonApprearFirstTime(String pattern,String person){
		String
personPattern=Pos.wildWord+Pos.wordDelimiter+Pos.NNP+"|"+"PERSON";
		if(pattern.equals(personPattern)){
			if(!personList.contains(person)){
				personList.add(person);
				this.isNameAppearFirstTime=true;
			}else{
				this.isNameAppearFirstTime=false;
			}
		}

	}



	/*this method will count how many verbs in each sentence
	 * in order to exclude some of the location patterns which
	 * might include a verb in the pattern
	 * @param token is the annotated sentence with word@Pos|NE
	 * @return number of verbs occurrance .
	 *
```

```
 * */
private int getNumOfVerbs(String token){
        int counter=0;
        Pattern pattern=Pattern.compile(Pos._VB);
        Matcher matcher=pattern.matcher(token);
        while(matcher.find()){
                counter++;
                //logger.info(matcher.group());
        }

        return counter;
}

/*this private method will parse the words from essay into
 * organised format
 * @param token words with POS, NE annotation
 * @return return the formatted string  */
private String printTuples(String token){
        String word=this.getMyWord(token);
        String pos=this.getMyPos(token);
        String ne=this.getMyNE(token);
        logger.debug(word+"\t|\t"+pos+"\t|\t"+ne);
```

```
            String str=word+"\t|\t"+pos+"\t|\t"+ne+"\n";


            return str;
    }


    private String getMyWord(String token){
            String[] words=token.split(Pos.wordDelimiter);
            if(words.length==0){
                    throw new IllegalArgumentException();
            }
            String word=words[0];
            return word;
    }


    private String getMyPos(String token){
            String[] words=token.split(Pos.wordDelimiter);
            if(words.length==0){
                    throw new IllegalArgumentException();
            }
            String[] posNE=words[1].trim().split(Pos.delimiter);
            String pos=posNE[0];
            return pos;
```

```
	}


	private String getMyNE(String token){
		String[] words=token.split(Pos.wordDelimiter);
		if(words.length==0){
			throw new IllegalArgumentException();
		}
		String[] posNE=words[1].trim().split(Pos.delimiter);
		logger.debug("|size is "+posNE.length+"  and words[1] is "+words[1]);
		logger.debug("PosNE 0 is "+posNE[0]+" POSNE1 is "+posNE[1]);
		String ne=posNE[1];
		return ne;
	}


	public String testMyPattern_InternalUse(String testStr,String pattern){
		String isFind=null;
		Pattern p=Pattern.compile(pattern);
		Matcher matcher=p.matcher(testStr);
		while(matcher.find()){
			logger.debug(matcher.group());
			isFind=matcher.group();
		}
```

```
                logger.info(testStr+" -- "+pattern+"=="+isFind);
                return isFind;
        }

        /* the Main testing block*/
        public static void main(String[] args) throws ClassCastException, IOException,
ClassNotFoundException{
                String
DT_JJ_NN=Pos.wildWord+Pos.wordDelimiter+Pos._DT+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._JJ+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._NN;
                String
DT_VB_NN=Pos.wildWord+Pos.wordDelimiter+Pos._DT+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._VB+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._NN;
                String
IN_DT_NN=Pos.wildWord+Pos.wordDelimiter+Pos._IN+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._DT+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._NN;
                String
IN_NN_NN=Pos.wildWord+Pos.wordDelimiter+Pos._IN+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._NN+Pos.connector+
```

```
                Pos.wildWord+Pos.wordDelimiter+Pos._NN;

                String
IN_PRP$_NN=Pos.wildWord+Pos.wordDelimiter+Pos._IN+Pos.connector+

                Pos.wildWord+Pos.wordDelimiter+Pos._PRP$+Pos.connector+

                Pos.wildWord+Pos.wordDelimiter+Pos._NN;

                String
NN_JJ_NN=Pos.wildWord+Pos.wordDelimiter+Pos._NN+Pos.connector+

                Pos.wildWord+Pos.wordDelimiter+Pos._JJ+Pos.connector+

                Pos.wildWord+Pos.wordDelimiter+Pos._NN;

                String
NN_DT_NN=Pos.wildWord+Pos.wordDelimiter+Pos._NN+Pos.connector+

                Pos.wildWord+Pos.wordDelimiter+Pos._DT+Pos.connector+

                Pos.wildWord+Pos.wordDelimiter+Pos._NN;

                String
NN_IN_NN=Pos.wildWord+Pos.wordDelimiter+Pos._NN+Pos.connector+

                Pos.wildWord+Pos.wordDelimiter+Pos._IN+Pos.connector+

                Pos.wildWord+Pos.wordDelimiter+Pos._NN;

                String
NN_IN_DT=Pos.wildWord+Pos.wordDelimiter+Pos._NN+Pos.connector+

                Pos.wildWord+Pos.wordDelimiter+Pos._IN+Pos.connector+

                Pos.wildWord+Pos.wordDelimiter+Pos._DT;

                String                                                  PRP$_JJ_NN=
```

```
        Pos.wildWord+Pos.wordDelimiter+Pos._PRP$+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._JJ+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._NN;
                String
RB_IN_NN=Pos.wildWord+Pos.wordDelimiter+Pos._RB+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._IN+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._NN;
                String
TO_DT_NN=Pos.wildWord+Pos.wordDelimiter+Pos._TO+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._DT+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._NN;
                String
VB_DT_NN=Pos.wildWord+Pos.wordDelimiter+Pos._VB+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._DT+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._NN;
                String
VB_PRP$_NN=Pos.wildWord+Pos.wordDelimiter+Pos._VB+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._PRP$+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._NN;
                String
VB_IN_NN=Pos.wildWord+Pos.wordDelimiter+Pos._VB+Pos.connector+
                Pos.wildWord+Pos.wordDelimiter+Pos._IN+Pos.connector+
```

```
                Pos.wildWord+Pos.wordDelimiter+Pos._NN;

                String
JJ_JJ_NN=Pos.wildWord+Pos.wordDelimiter+Pos._JJ+Pos.connector+

                Pos.wildWord+Pos.wordDelimiter+Pos._JJ+Pos.connector+

                Pos.wildWord+Pos.wordDelimiter+Pos._NN;


                String                                                  patternStr=
Pos.wildWord+Pos.wordDelimiter+Pos.NNP+Pos.delimiter+"PERSON"+Pos.or+

                Pos.wildWord+Pos.wordDelimiter+Pos._PRP+Pos.or+

                Pos.wildWord+Pos.wordDelimiter+Pos._NNP+Pos.tokenSplit+

                Pos.wildWord+Pos.wordDelimiter+Pos._VB+Pos.tokenSplit+

                DT_JJ_NN+Pos.or+

                DT_VB_NN+Pos.or+

                IN_DT_NN+Pos.or+

                //IN_NN_NN+Pos.or+

                IN_PRP$_NN+Pos.or+

                NN_JJ_NN+Pos.or+

                NN_DT_NN+Pos.or+

                NN_IN_NN+Pos.or+

                NN_IN_DT+Pos.or+

                PRP$_JJ_NN+Pos.or+

                RB_IN_NN+Pos.or+
```

```
TO_DT_NN+Pos.or+

VB_DT_NN+Pos.or+

VB_PRP$_NN+Pos.or+

VB_IN_NN+Pos.or+

JJ_JJ_NN+Pos.or+

Pos.wildWord+Pos.wordDelimiter+State.TIME+Pos.or+

Pos.wildWord+Pos.wordDelimiter+State.LOCATION;

Map<String,String> nerOverwrite=new HashMap<String,String>();



nerOverwrite.put(DT_JJ_NN, "Conditional Location");

nerOverwrite.put(IN_PRP$_NN, "Conditional Location");

nerOverwrite.put(IN_DT_NN, "Conditional Location");

nerOverwrite.put(RB_IN_NN, "Conditional Location");

nerOverwrite.put(TO_DT_NN, "Conditional Location");

nerOverwrite.put(VB_DT_NN, "Conditional Location");

nerOverwrite.put(VB_PRP$_NN, "Conditional Location");

nerOverwrite.put(IN_NN_NN, "Conditional State");

nerOverwrite.put(DT_VB_NN, "Conditional State");

nerOverwrite.put(NN_JJ_NN, "Conditional State");

nerOverwrite.put(NN_DT_NN, "Conditional State");

nerOverwrite.put(NN_IN_NN, "Conditional State");
```

```
nerOverwrite.put(NN_IN_DT, "Conditional State");
nerOverwrite.put(PRP$_JJ_NN, "Conditional State");
nerOverwrite.put(VB_IN_NN, "Conditional State");
nerOverwrite.put(JJ_JJ_NN, "Conditional State");
AnnotateEssays an=new AnnotateEssays();
ReadDocs doc=new ReadDocs();
boolean writeTofile=true;
//String rootURL="C:\\AEG\\Test Samples\\";
String rootURL=GetPath.getEssayRootPath();;
ArrayList<String> fileNames=new ArrayList<String>();
//this will get all the essayNames under the RootURL
Utilities.getAllEssaysNamesByRootDir(fileNames, new File(rootURL));
for(String fileName:fileNames){
an.resetStateIndicator();
String inputStr=doc.readDoc(rootURL+fileName);
// String inputStr="Anne tried the key and
...";
an.setFileName(fileName);
//String inputStr=doc.readText("D:\\AEG\\sean\\location.txt");
//String inputStr="One day after school, Anne stepped on
something.";
```

```
                    an.annotateEssays(inputStr,patternStr,nerOverwrite,writeTofile);

                }

        }

}
```

Appendix E - Event Detection Results

	Actual		Test	
No.	*Events*	*Non-Events*	*Events*	*Non-Events*
1	24	18	23	19
2	25	47	21	51
3	22	18	21	19
4	19	39	16	42
5	25	13	21	17
6	20	20	16	24
7	15	4	14	5
8	22	16	21	17
9	11	3	14	0
10	15	7	15	7
11	27	21	24	24
12	16	4	15	5
13	16	16	25	7
14	34	24	32	26
15	11	25	20	16
16	13	18	17	14
17	22	30	20	32
18	6	4	8	2
19	13	22	16	19
20	19	17	24	12
21	16	17	15	18
22	14	12	16	10
23	16	6	15	7
24	29	24	25	28
25	30	36	25	41
26	18	36	16	38
27	19	10	21	8
28	18	11	17	12
29	30	19	30	19
30	22	14	22	14
31	19	23	18	24
32	27	27	22	32
33	17	10	15	12
34	18	26	19	25
35	14	21	16	19
Total:	**682**	**658**	**675**	**665**

No.	*True Positives*	*True Negatives*	*False Positives*	*False Negatives*
1	20	15	3	4
2	19	45	2	6
3	19	16	2	3
4	14	39	0	5
5	21	13	0	4
6	12	20	0	8
7	13	3	1	2
8	17	12	4	5
9	11	0	3	0
10	12	5	2	3
11	21	18	3	6
12	14	3	2	1
13	16	7	9	0
14	25	16	8	9
15	10	15	10	1
16	13	14	4	0
17	17	27	3	5
18	6	2	2	0
19	13	18	4	0
20	18	10	7	1
21	13	15	2	3
22	14	10	2	0
23	13	4	2	3
24	24	23	1	5
25	24	36	0	6
26	13	33	3	5
27	17	6	4	2
28	16	10	1	2
29	27	17	2	3
30	19	11	3	3
31	17	22	1	2
32	19	24	3	8
33	14	10	0	3
34	15	21	5	3
35	12	16	5	2
Total:	**568**	**556**	**103**	**113**

Appendix F - Human marker assigned band scores

Last Name	*First Name*	*Audience*	*Ideas*	*Character Setting*	*Cohesion*
ADANO	KELLY	4	3	3	3
ADKINS	BRENT	2	2	2	2
AGENBAG	TREVOR	2	2	1	2
AMESS	LISA	4	4	3	3
ANDREWS	SHELBY	5	4	3	3
AZMI	NUR NADIA	4	3	2	2
BAGIATIS	ADELE	5	4	3	3
BAKER	CLAIRE	5	4	4	3
BAKER	LAURA	4	3	3	3
BEAVEN	PATRIC	2	2	1	2
BELLIS	JESSICA	5	5	4	4
BENNETT	KIMBERLY	3	3	2	2
BERENTE	JOSHUA	4	3	3	3
BERTOLA	CLAIRE	3	3	2	2
BETTI	EMMA	4	4	3	3
BIRSS	ELEANOR	4	3	2	3
BOCCAMAZZO	CAITLIN	5	4	3	3
BOCCAMAZZO	DAMIEN	4	3	3	3
BOLES-RYAN	AARON	5	5	3	3
BOTHMA	CORBAN	5	5	3	3
BOTH-WATSON	SERENA	5	5	3	4
BOWEN	CAITLIN	5	5	4	3
BRAMPTON	SHANI	4	4	3	4
BREAN	VERITY	5	4	3	4
BRIGGS	KYLE	2	1	1	2
BYRNES	WILLA	5	4	4	4
CABUNALDA	ANASTAJIA	5	4	3	3
CASTAING	JULIA	5	4	3	4
CATOVIC	NINA	5	3	4	4
CHANDLER	MORGAN	5	4	4	4
CHARLES	SINEAD	4	4	3	3
CHEDID	DANIELLE	6	4	4	4
CHEREL	ESMAY	2	1	1	1
CHETWYND	RHIANNA	2	2	1	2
CHU	WAIKEI	4	4	2	3
COLBY	HADDON	4	3	3	3
COMBI	WYATT	2	2	1	2
CONN	KRISTOPHER	4	3	2	2
COPPARD	MIIKA	5	4	4	4
COWELL	TANIEKA	5	4	4	3

COYNE	ROBERT	2	2	2	2
CUNNINGHAM	ANYA	4	4	3	3
DALE-FRASER	THOR	2	1	1	1
DANKS	LAUREN	5	4	4	4
DARCEY	PHILLIP	2	2	1	2
DE PLEDGE	HAYLEY	4	3	2	3
DIXON	JACOB	4	4	3	3
DOPOE	NEWON	2	2	2	2
DORAN	HAZEL	5	5	4	4
DORRELL	VICKY	4	3	2	3
ELLERTON	JAMIE	2	2	1	2
ESTENS	GEORGIE	4	4	3	2
FARLEY	TAYLOR	2	2	1	2
FARRELL	DAVID	2	1	1	1
FERGUSON	THOMAS	2	2	2	2
FOO	MELANIE	6	5	4	4
FORWARD	KYLE	5	4	4	3
GALANTE	MICHAEL	4	4	3	3
GESTE	IMOGEN	5	4	4	3
GIANATTI	ALEXANDRA	4	3	3	3
GORJY	DANIEL	4	3	2	3
GUTHRIE	DAVID	2	2	1	2
HAGUE	DYLAN	2	2	1	2
HAINES	EMMA	4	3	3	3
HALL	LEWIS	2	2	2	2
HANSEN	TRAE	3	2	2	2
HANSEN	RICHARD	2	1	0	1
HARLAND	JAYDEN	2	2	1	2
HAWKETT	MELISSA	5	4	3	3
HELSBY	EMMA	4	4	3	3
HENRY	JOSEPH	2	2	1	2
HIGGINSON	KAYLA	4	3	2	3
HODSON	KYLE	3	2	2	2
HOLT	LAUREN	4	3	3	3
HUDSON	BRYCE	2	2	2	2
HUNTER	BRADLEY	2	2	1	2
INGRAM	TOBY	4	4	3	3
IOPPOLO	CALEB	2	2	2	2
JOHNSON	RHYS	4	3	3	3
JONES	TYLER	2	2	2	2
KARSKI	TAHLIA	4	3	3	4
KELLY	DANIEL	3	3	2	2
KROLL	JASON	6	5	4	4
LOH	JILLIEN	6	5	4	4
MAIN	MARJORIE	6	5	4	4
MASON	NICHOLAS	3	3	2	3
MILTON	KYRON	2	2	2	2
PALAYUKAN	HONNY	6	5	4	3

PERSSON	MELANIE	6	5	4	4
VICKERY	MAXWELL	6	5	2	4

Appendix G - Precision, Recall, F-Measure and MCC values for Event Detection Process

No.	Precision	Recall	F-Measure	Matthew's Correlation Coefficient
1	0.87	0.83	0.85	0.54
2	0.9	0.76	0.83	0.43
3	0.9	0.86	0.88	0.58
4	1	0.74	0.85	0.42
5	1	0.84	0.91	0.63
6	1	0.6	0.75	0.4
7	0.93	0.87	0.9	0.53
8	0.81	0.77	0.79	0.44
9	0.79	1	0.88	0.48
10	0.86	0.8	0.83	0.45
11	0.88	0.78	0.82	0.49
12	0.88	0.93	0.9	0.56
13	0.64	1	0.78	0.55
14	0.76	0.74	0.75	0.36
15	0.5	0.91	0.65	0.42
16	0.76	1	0.87	0.61
17	0.85	0.77	0.81	0.46
18	0.75	1	0.86	0.61
19	0.76	1	0.87	0.59
20	0.72	0.95	0.82	0.55
21	0.87	0.81	0.84	0.51
22	0.88	1	0.93	0.7
23	0.87	0.81	0.84	0.43
24	0.96	0.83	0.89	0.57
25	1	0.8	0.89	0.52
26	0.81	0.72	0.76	0.39
27	0.81	0.89	0.85	0.5
28	0.94	0.89	0.91	0.64
29	0.93	0.9	0.92	0.64
30	0.86	0.86	0.86	0.56
31	0.94	0.89	0.92	0.58
32	0.86	0.7	0.78	0.43

33	1	0.82	0.9	0.61
34	0.75	0.83	0.79	0.47
35	0.71	0.86	0.77	0.47
Average:	0.85	0.85	0.84	0.52

Appendix H Chi Square Distribution Table

df	Probability of the Chi-Square [P (X^2)]								
	0.995	0.975	0.9	0.5	0.1	0.05	0.05	0.01	0.005
1	0.000	0.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860
5	0.412	0.831	1.610	4.351	0.236	11.070	12.832	15.086	16.750
6	0.676	1.237	2.402	5.348	10.645	12.592	14.449	16.812	18.548
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188

Appendix I - Score Comparison

Name	Audience	Ideas	C & S	Cohesion	Audience	Idea	C & S	Cohesion	Text structure	Vocabulary	Paragraph	Sentence Structure	Punctuation	Spelling	System	Human Marker
BIRSS	5.5	4.5	4	4	4	3	2	3	3	4	2	4	3	4	38	32
HANSEN	2	1.5	1.5	1.5	2	1	0	1	1	1	0	1	1	2	15.5	10
DE PLEDGE	5.5	4.5	3	4	4	3	2	3	3	3	2	3	3	4	35	30
BAKER	4	3	1.5	3	4	3	3	3	3	3	1	3	4	4	35.5	31
HAINES	5.5	4.5	4	3	4	3	3	3	3	4	2	4	3	4	37	33
HELSBY	5.5	4.5	4	4	4	4	3	3	3	3	2	4	3	4	37	33
HOLT	5.5	4.5	4	3	4	3	3	3	3	3	2	4	3	4	36	32
AMESS	5.5	4.5	3	4	4	4	3	3	3	3	1	4	3	5	36	33
GIANATTI	5.5	4.5	3	3	4	3	3	3	3	4	1	4	3	5	36	33
ADANO	5.5	4.5	3	3	4	3	3	3	3	3	2	4	3	3	34	31
BERENTE	5.5	4.5	3	3	4	3	3	3	1	3	1	4	3	4	32	29
CHU	4	4.5	3	4	4	4	2	3	3	3	2	4	4	4	35.5	33
DORRELL	5.5	3	3	3	4	3	2	3	3	3	2	4	3	4	33.5	31
HIGGINSON	4	4.5	3	3	4	3	2	3	3	3	2	3	3	3	31.5	29
BAGIATIS	5.5	4.5	4	3	5	4	3	3	3	4	2	5	5	5	41	39
COWELL	5.5	4.5	4	4	5	4	4	3	4	4	1	4	3	4	38	36
BOCCAMAZZO	4	3	3	4	4	3	3	3	3	3	0	4	4	5	34	32
CONN	4	3	3	3	4	3	2	2	2	4	1	4	3	4	31	29
KARSKI	4	4.5	3	4	4	3	3	4	3	4	2	5	5	5	39.5	38
DALE-FRASER	2	1.5	1.5	1.5	2	1	1	1	1	2	0	2	1	2	14.5	13
CHEREL	2	1.5	1.5	1.5	2	1	1	1	2	2	0	0	0	2	12.5	11
FARRELL	2	1.5	1.5	1.5	2	1	1	1	1	1	0	1	1	2	12.5	11
BOTH-WATSON	5.5	4.5	4	4	5	5	3	4	4	5	2	5	5	6	45	44
CASTAING	5.5	4.5	4	3	5	4	3	4	4	4	2	5	3	4	39	38
COPPARD	5.5	4.5	4	4	5	4	4	4	4	4	2	4	4	3	39	38

CABUNALDA	5.5	4.5	3	3	5	4	3	3	3	4	2	4	4	5	38	37
HAWKETT	5.5	4.5	3	3	5	4	3	3	3	4	2	5	4	4	38	37
JOHNSON	4	3	3	4	4	3	3	3	3	4	1	4	2	4	32	31
AGENBAG	2	3	1.5	1.5	2	2	1	2	2	2	1	2	2	2	19	18
BOLES-RYAN	5.5	3	4	4	5	5	3	3	4	5	1	6	5	5	42.5	42
BRAMPTON	5.5	3	3	4	4	4	3	4	3	4	0	4	4	6	36.5	36
DIXON	5.5	3	3	3	4	4	3	3	3	4	2	4	4	4	35.5	35
BETTI	4	4.5	3	3	4	4	3	3	3	3	1	4	3	3	31.5	31
BRIGGS	2	1.5	1.5	1.5	2	1	1	2	2	2	0	1	0	2	13.5	13
PALAYUKAN	5.5	4.5	4	4	6	5	4	3	4	4	2	6	5	5	44	44
GESTE	5.5	4.5	3	3	5	4	4	3	3	4	2	4	3	4	36	36
BOCCAMAZZO	5.5	4.5	3	3	5	4	3	3	3	3	2	4	4	4	35	35
ESTENS	4	3	3	3	4	4	3	2	3	3	1	3	3	4	30	30
BREAN	4	4.5	3	4	5	4	3	4	4	4	2	5	4	5	39.5	40
ANDREWS	5.5	3	3	3	5	4	3	3	3	4	2	4	4	4	35.5	36
BERTOLA	2	1.5	3	3	3	3	2	2	2	3	1	3	3	4	25.5	26
DARCEY	2	1.5	1.5	1.5	2	2	1	2	2	2	0	3	3	3	19.5	20
BEAVEN	2	1.5	1.5	1.5	2	2	1	2	2	2	1	2	2	3	18.5	19
COMBI	2	1.5	1.5	1.5	2	2	1	2	2	2	0	2	2	3	17.5	18
CHETWYND	2	1.5	1.5	1.5	2	2	1	2	1	2	0	2	2	3	16.5	17
FARLEY	2	1.5	1.5	1.5	2	2	1	2	2	2	0	2	1	3	16.5	17
GUTHRIE	2	1.5	1.5	1.5	2	2	1	2	2	2	0	2	1	2	15.5	16
HARLAND	2	1.5	1.5	1.5	2	2	1	2	1	2	0	3	1	2	15.5	16
HUNTER	2	1.5	1.5	1.5	2	2	1	2	1	2	1	2	1	2	15.5	16
ELLERTON	2	1.5	1.5	1.5	2	2	1	2	2	2	0	1	1	2	14.5	15
HENRY	2	1.5	1.5	1.5	2	2	1	2	1	2	0	2	1	2	14.5	15
HAGUE	2	1.5	1.5	1.5	2	2	1	2	1	2	0	1	0	2	12.5	13
PERSSON	5.5	4.5	4	4	6	5	4	4	4	5	2	5	5	6	45	46
KROLL	5.5	4.5	4	4	6	5	4	4	3	5	2	6	4	6	44	45
LOH	5.5	4.5	4	4	6	5	4	4	4	4	2	6	5	5	44	45
CHEDID	5.5	4.5	4	3	6	4	4	4	3	5	0	5	4	6	40	41
VICKERY	5.5	4.5	3	3	6	5	2	4	4	5	2	5	3	5	40	41

GALANTE	4	3	3	3	4	4	3	3	3	4	2	5	5	5	37	38
CHARLES	4	3	3	3	4	4	3	3	3	4	0	4	4	5	33	34
CUNNINGHAM	4	3	3	3	4	4	3	3	3	3	2	4	4	4	33	34
BOTHMA	5.5	3	3	3	5	5	3	3	4	4	2	5	4	4	37.5	39
BYRNES	4	4.5	3	4	5	4	4	4	3	4	1	5	4	5	37.5	39
DANKS	5.5	3	3	4	5	4	4	4	4	4	1	5	2	5	36.5	38
FORWARD	4	4.5	3	3	5	4	4	3	3	4	1	4	4	5	35.5	37
ADKINS	2	1.5	1.5	1.5	2	2	2	2	2	3	1	2	2	3	19.5	21
FERGUSON	2	1.5	1.5	1.5	2	2	2	2	2	2	0	3	2	3	18.5	20
JONES	2	1.5	1.5	1.5	2	2	2	2	2	2	1	2	3	2	18.5	20
MILTON	2	1.5	1.5	1.5	2	2	2	2	2	2	0	2	2	3	17.5	19
HALL	2	1.5	1.5	1.5	2	2	2	2	1	2	1	2	2	2	16.5	18
HUDSON	2	1.5	1.5	1.5	2	2	2	2	1	2	0	2	2	3	16.5	18
COYNE	2	1.5	1.5	1.5	2	2	2	2	1	2	0	2	1	3	15.5	17
DOPOE	2	1.5	1.5	1.5	2	2	2	2	1	2	0	2	2	2	15.5	17
IOPPOLO	2	1.5	1.5	1.5	2	2	2	2	1	2	0	2	2	2	15.5	17
MAIN	5.5	4.5	3	4	6	5	4	4	4	5	1	6	5	6	44	46
DORAN	5.5	4.5	3	3	5	5	4	4	3	5	2	5	4	4	39	41
BENNETT	2	1.5	1.5	3	3	3	2	2	3	3	2	3	2	4	25	27
FOO	5.5	3	4	4	6	5	4	4	4	5	2	6	4	5	42.5	45
GORJY	2	3	1.5	3	4	3	2	3	2	4	1	4	4	4	28.5	31
HODSON	2	1.5	1.5	1.5	3	2	2	2	2	2	0	2	1	3	16.5	19
CATOVIC	4	3	3	3	5	3	4	4	3	4	1	5	4	5	35	38
MASON	2	3	1.5	1.5	3	3	2	3	3	2	2	3	3	4	25	28
KELLY	2	1.5	1.5	1.5	3	3	2	2	2	2	1	3	2	3	19.5	23
BAKER	5.5	4.5	4	4	5	4	4	3	4	4	2	5	4	5	36	40
BOWEN	4	3	3	3	5	5	4	3	3	4	1	5	4	5	35	39
INGRAM	2	3	1.5	3	4	4	3	3	3	3	2	4	2	3	26.5	31
AZMI	2	1.5	1.5	1.5	4	3	2	2	2	3	1	3	3	4	22.5	27
BELLIS	2	3	4	4	5	5	4	4	3	4	1	5	4	5	35	40
COLBY	2	1.5	1.5	3	4	3	3	3	3	3	1	4	3	4	26	31
CHANDLER	4	3	1.5	3	5	4	4	4	3	4	2	5	4	5	34.5	40

HANSEN	2	1.5	1.5	1.5	3	2	2	2	1	2	0	2	1	3	12.5	18