Automated Essay Marking –
for both Style and Content

James R Christie
School of Computer and Mathematical Sciences
The Robert Gordon University, Aberdeen

Abstract

This paper covers the automated assessment of essays. Such marking requires the assessment of style, and where appropriate the assessment of content.

The first part of this paper describes the assessment of style; while the second part proceeds to describe the assessment of content.

Style

The methodology used to assess style is based on a set of 'common' metrics, and requires some initial calibration.

After briefly outlining the method of assessing style, this part concludes by posing these questions on some varied aspects on the use of a common metric set, namely:

How valid is the computer marking of style?
Could there be a standard metric set for common use?
What constitutes a standard or an optimal metric set to use?
What is the effect of the size of the calibration sample used?

Content

After the terms "usage" and "coverage" are defined one particular essay set is examined in detail. Detail in terms of usage and coverage will all be discussed.

The schema for content does not require extensive development before the commencement of the assessment of content. No calibration is required for the methodology used to mark content, although the usual practise of taking a sample to verify the method would be recommended.

Conclusion

At the current stage of this development the methodology provides an expandable, flexible method for the marking of the essay content; and also provides a method for the marking of the essay style; both markings being produced by the computer.

The essay set, that was used as a demonstration vehicle, will be used further to indicate the potential throughput that computer based marking may achieve.

Introduction

This paper covers the automated assessment of essays based on the author's ongoing PhD research work and the resulting software system called **SEAR**.

SEAR, the acronym, stands for **S**chema, **E**xtract, **A**nalyse and **R**eport.
The research work is about mid-stage towards the completion of the PhD so this paper could be viewed as a progress report following on from an earlier paper made by this author (Christie, 1998).

For a 'reasonable' definition of an essay the author directs you to the Stalnaker definition (Stalnaker, 1951).

The automated assessment of essays requires the assessment of style, and where appropriate the assessment of content. The first part of this paper describes the assessment of style; while the second part proceeds to describe the assessment of content.

Style

Virtually all essays are candidates for being marked for style - in fact it is difficult to envisage an essay that would not be suitable for style assessment! Some markers, however, may choose not to mark style!

SEAR is currently set up for using a fixed set of 'common' metrics as used by others (exampled by Page 1996, Slotnick 1972). Published papers, produced by other researchers, show that this approach is successful in that the performance of computer based marking is to worse than the performance produced by two human markers working in combination.

In essence for the computer marking of style the method is:

```
pre-determining what would be candidate metrics [the pool of metrics]

take a subset of essays for the essay set
    and mark them manually

process this subset by computer [calibration]
    adjusting the weight of each metric
        until an acceptable agreement
            between human and computer marking is achieved
```

processing of the whole essay set.

In other words the approach adopted by researchers to the computer marking style is to use a weighted linear function. This commonality of approach is both reassuring and worrying - reassuring, in that although different metrics are used the performance is the same; worrying, in that there appears to be an absence of alternative(s) to the use of metrics.

Moving onwards, this approach appears only to be / been applied to essays written in plain ASCII text.

This artificial limitation in the application of the approach ignores the effect(s) on the marking of style enabled by text enhancements, font [type, size], colour, paragraph / page formatting and inclusion of non-textual elements. Thus many style effects produced by modern word-processors is wasted.

To be pedantic most modern word-processor packages offer the user some readability statistics using measures based on 'classical' methods. The author himself produced software to measure the Modified Fog Index to help students with their essays using plain text documents (Christie, 1987 unpublished).

Points-to-ponder in assessing style via metrics

How valid is the computer marking of style?

The research already published by others indicates that it's valid!
So, perhaps the question should be "How acceptable, to humans, is the computer marking of style?". That's not so clearly answerable. There could be considerable resistance to the adoption of computer based marking of style.

Could there be a standard metric set for common use?

The more metrics used then the better the agreement with human markers should become, in theory.

But: There is always going to be a presence of disagreement between human markers,
so why 'force' the computer to compete with, or eliminate, this disagreement?

Usually in the workplace there is little chance of having double marking of essays,
and the computer is better than two markers!

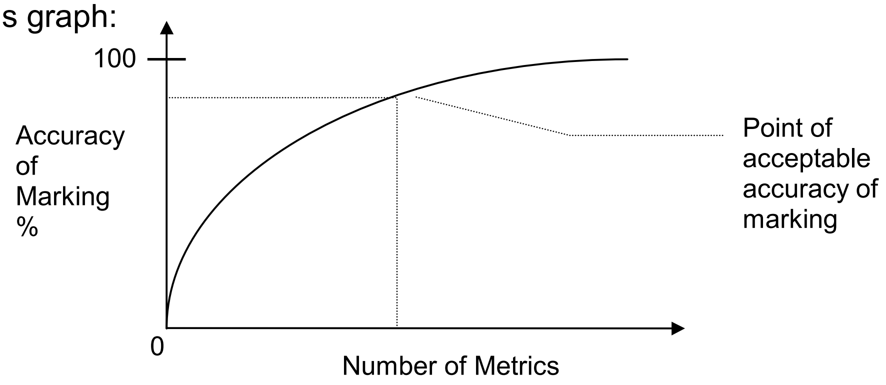
More metrics used,

the more processing time is required per essay,
the larger the calibration sample has to be,
the clearer the frequent use of some metric(s) over others would appear,

leading, back again, to the creation of a set of common metric(s).

What constitutes a standard, or an optimal, metric set to use?

Consider this graph:



The more metrics used the better the marking should become, as stated before.

One valid question is what number of metrics would result in perfect marking? Probably there will never be enough metrics available.

Devising more and more metrics would ultimately lead to the creation of ‘less’ sensible metrics - say “The percentage of five lettered words starting with the letter ’e’ used as the second word in a sentence”.

Equally valid is the question of how many metrics would give an acceptably accurate result?
This answer of acceptability sets the upper figure of metrics to use, whatever it is [about fifteen to twenty? (Christie, 1998)]. The specific metrics that should be used in a standard set are those that most frequently occur out of all the metrics devised so far, and those yet to be devised [see Conclusion below].

If it proves difficult to devise one standard set of metrics for common use, then the next best would be a suite of standard sets - one set used for a particular essay type.

Once determined then the standard set(s) would evolve with time, but slowly.

What is the effect of the size of the calibration sample used?

The more metrics used the bigger the sample for calibration has to become. To be statistically valid then a minimum sample size is twice the number of metrics to be used.
If you use a large number of metrics then it could arise that once you have marked the calibration sample, then there may be so few unmarked essays left - so why bother using the computer then?

As an aside the author has frequently heard the comment that it would be ‘un-safe’ to let the students know the metrics they will be marked against - why ‘un-safe’?

Surely knowing what the metrics to be used are is just as valid as the students knowing word count, number of pages, style of essay expected, points-to-ponder, guidance on spelling and grammar, and items to include / exclude in their essay.

Content

For content assessment then only those essays that are technical are candidates for this type of marking. An essay on “What five people would you choose for joining you in a dinner party, and why?” could lead to a vast spectrum of content, whereas an essay on “Describe the planetary motions within the Solar System” should lead to a bounded spectrum of content!

Interestingly, or is it worryingly, there appears to be a range of different approaches to the marking of content by computer. All the approaches appear to be as successful as the marking of style is. Some approaches require the computer to be ‘trained’ before it is set to mark the essays. These approaches may confer a degree of inflexibility in the essay set(s) that may be processed.

The author compounds this range of approaches by devising yet another one. For SEAR the content schema is prepared once [, and be revised] fairly quickly and easily. Further the SEAR content schema requires neither ‘training’ nor ‘calibration’. The schema is held as a simple data structure [that would not be out of place in an introductory COBOL programming class!]. As the author’s research is progressing it is appearing that the size of the data structure is likely to behave linearly with the size the schema.

The author has devised two measures to aid the marking content by computer - *usage* and *coverage*. In both measures high is good, and low is bad; although the interpretation of these is not the same. These measures were devised to envision the relationship between each essay and the schema.

Usage

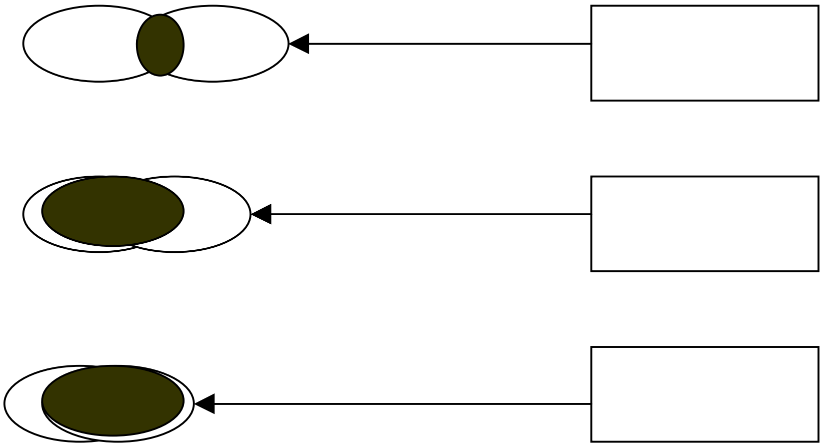
This is a measure of how much of each essay is used.
Consider an essay largely composed of bullet points. This essay would be succinct [and it would also have a poor style!] and to the point the usage will be high. On the other hand an essay scoring low on usage is potentially off the point or too wordy or is a bad faith essay.

Coverage

This is measure of how much of the schema is ‘used’ by the essay under examination.
High coverage [i.e. high content mark] means that the essay has contained all the items of the schema, with the relationship(s) between the items. Low coverage mark means the essay was potentially off the point or bad faith or something else.

An essay scoring high in both usage and coverage would represent a highly crafted essay.
An essay scoring both low should give the essay setter / human overseer causes for concern - either the essayist is struggling to achieve, or the essayist has produced a superior essay above and beyond the expectations [or capabilities] for the essay setter.

These diagrams indicate the interaction between coverage and usage:



One particular essay set

The author is indebted to Dr Butler for producing an essay set of five essays for to develop the initial data structure. These essays are derived from one of The

Robert Gordon University web pages (RGU, 1999) that provides a potted history of the University' founder Robert Gordon. These are the same essays used in the Summer 1999 Inter-Marker Survey.

From using this essay set, SEAR initially produced these figures:

Essay	Comment	Word Size	Coverage %	Usage %	Mark %
very good	good style	180	64	51	64
very good	poor style	264	56	30	56
good	Doubled up in size	520	64	35	64
poor	poor style	275	57	27	57
'bad faith'	to deceive the software	105	18	21	18

The coverage and mark agree, as expected; the word count is correct while the usage is about right.

However there is some misalignment of the coverage. The coverage is lower than expected for the better essays, yet higher than expected for the poorer essays. The reason for this is the manner of handling the relationships between the various items of the schema.

Given that these are the initial values produced subsequent development versions of SEAR software better coverage % have been achieved.

Conclusion

At the current stage of this research the SEAR methodology provides an expandable, flexible method for the marking of the essay content; and also provides a method for the marking of the essay style; both markings being produced by the computer.

For style the classical metrics are to be augmented by inclusion of metrics that relate to word-processing functionality for the release version of SEAR. There is nothing to suggest that the inclusion of this genre of metrics will produce any adverse effect(s) on marking. This will increase the scope of metrics that may be used for a common set that can by applied in style marking. Removing the restriction of a common set of metrics would lead to a system of selecting x metrics from y options for each particular essay set - more statistic processing will be required by the marker, however that is not insurmountable and can be automated (Page, 1996).

For content the data structure currently being developed as SEAR has the potential to cope with very large content schema without the need for any training sets. The author would commend that samples from the essay set are tested before the complete essay set is processed.

Although student and staff feedback is provided by SEAR at the moment, field trialing from the summer of 1999 onwards is expected to refine and enhance the feedback functionality.

Throughput

Currently SEAR operates on English language documents produced by various versions of MS Word™. The software is being developed in the C/C++ programming language on a 486-66 PC with 16MB RAM, 1GB Hard Drive using Win95™ for the operating system.

The Schema and Report parts of SEAR are used as needed.
As does the processing of calibration sample for style.
The Extract and Analyse parts process the essay set.

The operating system used affects the size of the word-processed document produced by the various versions of Word™ basing an essay create using Word™ 6 on Windows 3.1™ scaled as 100, then the difference in file size by version and operating system is:

Operating System	**Word™ Version**		
	6 [MS Office 4.3™]	**7 [MS Office 95™]**	**97 [MS Office 97™]**
Windows 3.1™	100	n/a	n/a
Windows 95™	230	160	270
Windows NT™	n/a	n/a	270

In terms of throughput the table below indicate the timings currently being achieved for multiple copies of the same essay of ~300 words produced by different versions of Word™ running on the hardware specified for the SEAR software:

Function	Time for 500 essays	Rate per hour [essays per hour]
Extract		
Word 6™	104 seconds	~17,000
Word 7™	106 seconds	~17,000
Word 97™	120 seconds	~15,000
Analysis		
Style	42 minutes	~700
Content	to be announced	at conference

The difference in the extraction rates for the various versions of Word™ are due to the differences in file size when documents are produced. For a given essay the extracted files are the same regardless of which word-processor and which operating system were used to create the original essay.

The analysis [style, content or both] is conducted using the extracted files.

In summary throughput depends on:

original essay size - essay size in words [obviously],
the version of word-processor used,
and operating system used;

number of essays in the essay set;

the level of functionality of the Extract and Analysis software:
increasing functionality will decrease throughput;

using faster hardware and operating systems will increase throughput;
using a [busy] network will decrease throughput.

Future expansion beyond the completion of the author's PhD

In the long term SEAR has the potential to:

operate on word-processed documents produced using other software;
provide measures for subjectivity, opinion, tension, etc.,
offer plagiarism detection.

In the **very** long term SEAR has the potential to:

operate with non-English languages.

Acknowledgements

The author acknowledges the support, enthusiasm and patience shown by his PhD supervisory team:

Prof. H Ellington, Director of Studies [Robert Gordon University],
Dr R Butler [Robert Gordon University] and
Mrs S Earl [Napier University]

References

Christie, J. R. (1987) Modified Fog Index. Unpublished material.
Software for students to use to improve the readability of their essays.

Christie, J. R. (1998) ‘Computer-assisted Assessment of Essays’,
Proceedings of the Second Annual Computer Assisted Assessment Conference 1998, The Flexible Learning Initiative, Loughborough University, p 85 - 89

Page, E. B. et al (1996) ‘Computer Grading of Essay Traits in Student Writing’
NCME, New York, p 1 - 8

Slotnick, H. B. (1972) ‘Toward a Theory of Computer Essay Grading’
Journal of Educational Measurement, vol 9, no 4, p 253 - 263

Stalnaker (1951) Quoted by Coffman, W. E. ‘Essay Examination’.
In: R. L. Thorndike Educational Measurement 2nd edition Chapter 10, 271.

RGU (1999) http://www.rgu.ac.uk/topic/theuni/robg.htm
The Robert Gordon University web page, as at Monday 26th April 1999.
This web page describes the history of Robert Gordon himself, and the founding of the University that bears his name.