EXPLORING THE USE OF COMPUTATIONAL LINGUISTICS FOR AUTOMATED FORMATIVE FEEDBACK IN THE HUMANITIES[1]

Jessie Paterson[1], Christian Lange[1], Iqbal Akhtar[1], Francisco Iacobelli[2], Paul Anderson[3], Annette Leonhard[3]

[1]School of Divinity, University of Edinburgh, UK
[2]Northwestern University, USA
[3]School of Informatics, University of Edinburgh, UK
jessie.paterson@ed.ac.uk, c.lange@ed.ac.uk, s0880824@sms.ed.ac.uk, franciscoiacobelli2011@u.northwestern.edu, dcspaul@ed.ac.uk, annette.leonhard@ed.ac.uk

Abstract

This paper presents the results an investigation into the possible role of computational linguistic techniques in providing automated formative feedback on student's written work - including both traditional essays, and collaborative wikis.

We first attempt to identify some of the criteria used by academic staff when marking student work. Using real examples from the School of Divinity at Edinburgh University, we analyse written feedback from the markers to produce an explicit list of criteria. We also use a number of automated tools to analyse certain features of the work and identify those which correlate with the perceived "quality".

We then survey various techniques from computational linguistics to determine how they might potentially be used to identify some of these criteria automatically.

We conclude that there is a real potential to produce an automated tool which would provide practically useful formative feedback. We also note that the process of rigorously defining the criteria was helpful to the academic staff in clarifying their manual marking process.

Keywords: automated formative feedback, computational linguistics

1 INTRODUCTION

Marking essay-type work is often controlled by rigid marking criteria and quality control procedures, such as the "Common Marking Scheme" used at the University of Edinburgh. However, this still relies on their interpretation by individual markers who apply their own implicit criteria. For example, when we began to look more closely at the marking of wikis and traditional essays, it became clear that the quality criteria being used were distinctly different, although the markers had often not been conscious of this. This led to an interest in how we might help students to understand these criteria, and their relationship to the written exercise before submitting the work.

The literature has numerous examples of the importance of such "formative" feedback in the whole assessment process. Black and William [1] provide quite a rigorous review of the literature (for the time). This provides substantial evidence that "innovations designed to strengthen the frequent feedback that students receive about their learning yield substantial learning gains". Boud [2] strongly emphasised the role of formative assessment and Knight [3] proposed that summative assessment is in disarray and there is more need to emphasis formative assessment. More recently, frameworks for assessment such as those proposed by Carless [4], Gibbs and Simpson [5] and Nicol and MacFarlane-Dick [6] have further explored the role of formative feedback. The work of Nicol has been the basis for *The Re-Engineering Assessment Practices in Scottish Higher Education* project[2] where feedback is seen as a part of the student's role in regulating learning. The *Enhancement themes* have produced a series of resources on Integrative assessment[3] including the publication on "Balancing

[1] This project has been funded by IDEA lab: http://forum.idea.ed.ac.uk
[2] http://www.reap.ac.uk/resourcesPrinciples.html
[3] http://www.enhancementthemes.ac.uk/publications/default.asp#Integrative

Assessment of and Assessment for Learning"[4] which makes recommendations about the importance of feedback (or more correctly "feedforward"). This is further explored in the TLA newsletter by Hounsell [7] in which he explores further the ideas behind feedforward and its relationship to the student experience. Encouragingly it has been reported that students find formative feedback provided electronically as useful as that provided by markers [8].

In order to enhance the student feedback, we wanted to see if we could use techniques from computational linguistics to create a tool which would provide formative feedback and guidance on the quality criteria which were actually being applying during marking. Section 2 describes our investigation into the quality criteria and attempts to describe these in an explicit way. Section 3 surveys various computational techniques and discusses how they may be used to automatically identify some of these criteria.

2 DEFINING THE QUALITY CRITERIA

To explore the quality criteria actually being used in practical marking, we used real data and examples from one course - "Martyrs, virgins, and hell's angels: Islamic eschatology in context", a level 10 course taught over one semester in the School of Divinity, at the University of Edinburgh. There were approximately thirty students in the class and each year they were asked to produce a traditional essay as well as contribute to a collaborative wiki. For the wiki, each student is allocated a subject and asked to work together with two other colleagues. They are asked to write 1000 words each towards a 3000 word wiki. Each students' work is clearly differentiated, and the students are assessed on their own work as well as the final joint effort. Data from 2008-09 and 2009-10 academic years was available. For 2008-09 paper copies of the wikis with written comments were available but only the uncommented essays were available. For 2009-10, electronic wiki comments, and the written comments on the essays were available. The marks for all the components were also available. This raw data formed the basis of the project.

Intuitively, it seems that likely that many of these findings will be applicable to a wide range of subject matter, but they have been determined with reference to one specific course, and some caution should be exercised when generalising.

2.1 Observed Criteria

The following quality criteria were determined by examining the written work and comparing the awarded marks with the comments provided by the marker to extract the features that differentiated the work. These were refined through discussions with the marker to produce a list of criteria applying both to the wiki entries and the essays:

2.1.1 Referencing

- The use of primary sources in referencing is highly encouraged.
- Peer reviewed academic print journals are generally considered better than the use of solely electronic resources.
- Use of more recent publications provides the latest ideas in a specific field is, though it is acknowledged that every field has its classics.
- Consistence in citing references is a positive feature.
- When citing directly from sources, use appropriate quotations and citations. Overuse of quotations is discouraged and generally no more than about 15% of an essay should consist of direct quotations.
- Use many different sources in referencing appropriately.

[4] http://www.enhancementthemes.ac.uk/documents/IntegrativeAssessment/IA%20Balancing%20assessment.pdf

2.1.2 Style/Terminology

- Vary the writing and use precise wording. Terms such as “illuminate”, “criticism”, and “discourse” are more precise in meaning than simple verbs or nouns. Structurally complex sentences help to provide variation.
- Use the active voice and abstain from the passive voice as much as possible. Run-on sentences should be separated into two sentences.
- The use of unique nouns and verbs helps to convey the subtlety of arguments. Avoid colloquialisms such as “hugely” and “massively”. Vague terms such as “interesting” and “nice” should also be avoided.
- Avoid making normative claims.

2.1.3 Structure

- The first paragraph of our work should outline the main argument. The conclusion should be a summary of the intervening sections and final thoughts. There ought to be a similarity in the terminology used in the introduction and conclusion.
- When appropriate, employ discursive writing techniques.
- Have an idea of your overall argument and the word limit.
- Present arguments clearly. Avoid obscure writing.

2.2 Automated Analysis

In addition to the criteria identified manually, we performed some automated analysis of the student work to see if this could highlight further features. We used two different tools to perform this analysis: *Linguistic Inquiry and Word Count* (LIWC), and *WordSmith Tools*[5].

We chose a selection of high and low graded work and compared this to the overall essays or wikis. The statistically significant differences identified by LIWC were:-

For the wiki:

Grammar - More use of “I” and “we”, less use of “they”; more articles and auxiliary verbs; more past and future tense, less present tense; fewer adjectives that negate; less use of “affect”; less use of negative emotive words; less use of words that reflect anger; more use of words dealing with perceptions; less use of “feel”, more use of “time” words such as dates; less use of the “religion” terms.

Punctuation - Less use of the period, semicolon, apostrophe, parentheses, and punctuation marks overall; more use of comma, colon, question mark, dash, and quotation marks (inverted commas).

For the essays:

Grammar - Higher word count; less use of pronouns, higher use of proper nouns; less use of auxiliary verbs; less past and present tense used in favour of future tense; less use of negation; more use of terminology related to social organisation; less use of words relating to emotion, (positive or negative); less terminology related to theorising causation; less use of sense verbs (feel, see, etc...); more use of words related to religion.

Punctuation - More use of the period, colon, question mark, dash, quotation marks (inverted commas), parentheses, and punctuation marks overall; less use of semi-colon and apostrophe

The Wordsmith tools showed that higher marked essays:

Have more distinct terms (non-repetitive); more higher lettered words; longer vocabulary terms; lower type/token ratio; more words per sentence on average; more proper nouns; less use of prepositions and conjunctive words on average.

Both techniques identified a number of trends that had not been identified by the manual process. This suggests that mixture of both manual and automatic processes should be used when attempting to define the criteria.

---

[5] http://www.liwc.net/, http://www.lexically.net/wordsmith/

3 POTENTIAL TECHNIQUES

The previous section identified some of the criteria which we consider to be important in the student essays and wikis. In this section, we survey a range of techniques from computational linguistics to assess the potential for providing automatic feedback on these aspects. In general, these techniques can be divided into those which attempt to analyse the "content" of the text (section 3.2), and those which perform a more "surface analysis" - references, grammar, etc. (section 3.3). Both of these are valuable in this context, however some approaches to content analysis rely on a large body of sample text for "training", and this is not (currently) available for our application.

There is a considerable body of work on automated assessment (see [9] for one survey). Clearly, many of these techniques are also appropriate for providing student feedback. However, tentative results and suggestions are often valuable as feedback when they would not be appropriate for use in assessment. This means that a wider range of techniques may be practical in our application.

3.1 Background

Researchers have been concerned with tools to help good writing style for some time now. For example, Kiefer and colleagues [10] produced a text analyser called Writer's Workbench[6] which is now commercial software. Kieras [11] developed a system to help the U.S. Navy with writing style on technical reports. More recently, Select-a-Kibitzer [12] and its close relative, StoryStation [13] used Latent Semantic Analysis (LSA) [14] to determine cohesion between sentences and provide feedback on the composition.

There are other variations of semantic analysis for essays, such as PEG and the PEG-based "E-rater" [15]. These grade essays using LSA and regression measures, and consider word usage, readability and a few other heuristics to provide feedback to the writer. E-rater does not currently consider a wide range of surface features related to writing style -in particular those that have to do with the language of research essays specified in Section 2. E-rater, like all approaches using LSA, also needs extensive training with previous essays. Tristan Miller [16] provides a good discussion with evaluations of these systems, although the E-rater discussion is somewhat outdated.

All of these approaches are capable of providing users with feedback on their writing style. However, this feedback is not always very specific and does not address the inherent structure of an academic essay or Wiki entry. In particular, they do not take into account format, quantity and reliability of sources used. Neither do they provide metrics to assess whether an essay is presenting a coherent question, or using terms consistently throughout the text (for example "Islam" and "Muslim").

Halliday and Hasan [17] analysed linguistic features of text in English and formalised the concepts of *cohesion* and *coherence*. Meurer [18] found that good essays used significantly more lexical cohesion ties than referential ties. Bae [19] found that lexical cohesion and local references correlated highly with coherence in children's essays. While cohesive measures are relatively systematic and easy to identify [20], coherence measures are more subjective. However, proxies for cohesion have been successfully computed in some cases [21] using simple measures of lexical cohesiveness.

Pennebaker and King [22] carried out an analysis of essays using LIWC and found 15 kinds of words that correlated highly (positively or negatively) with four factors they termed "immediacy," "making distinctions," "the social past" and "rationalization."

Finally, Graesser, McNamara and Louwerse [23] offer a very good and comprehensive review on the current metrics and tools available to analyse text automatically.

3.2 Content Feedback

The goals and techniques presented in this section aim to detect potential issues related to content and its presentation in essays. These rely on cognitive assumptions - i.e. the metrics described are reasonable indicators of the thought process of the authors while producing their essays. Other techniques to detect surface features with no underlying assumptions are described in the following section.

---

[6] http://www.emo.com/wwb/wwbEnhancements.htm

3.2.1 Clear Question and Thesis

It seems that there are markers in the first paragraph that can be good a proxy to determine whether the introduction is motivating, and poses a question and a thesis. In addition, analogous measures of text similarity can be used to check for lexical cohesion between the introduction and conclusion. This provides a way of assessing compliance with rule 1 from section 2.1.3.

• TextTiling [24], perhaps enhanced with LSA [25], can be used to examine the introduction of the essay. TextTiling can be used to assess inter-sentence similarity (an indicator of lexical cohesion in text) and the similarity between the title of the essay and the last sentence of the introduction. The intuition behind this metric is that a cohesive (generally well written) introduction that ends with a question similar to the title is probably an introduction that makes sense and that smoothly (cohesively) leads into the question posed by the title. Section 4.1 shows the results of some experimental analysis which support this hypothesis.

3.2.2 Sufficient Context

"Sufficient context" refers to a balance among theological, social, historical and anthropological context where it is due. This is somewhat subjective, but the following extract is from an essay which had been given this comment by the marker:

Zoroastrian acceptance to heaven or hell begins with the judgment of the soul based (Y,31:14). This takes place on the 'Chinvat Bridge' or 'Bridge of Seperator' that leads to heaven (AVN, ch.2). This notion is comparable to that of the Al-Sirat in the Hadiths, where the Day of Judgment is described as passing over Hell on a narrow bridge in order to enter Paradise. However this will occur on the Yawm al-Qiyāmah (Day of Judgment) whereas in Zoroastrianism it occurs before Frashegird (Day of Judgment) suggesting outside influence.

Besides the fact that the first sentence is incomplete, there seems to be an uncomfortable leap to suggest an outside influence on the theology about judgment day. Perhaps a better paragraph would have included historic events or evidence of Islam adapting other ideas from Zoroastrianism. Perhaps by including intermediate steps that helped preserve certain beliefs, the influence can be more strongly established.

Computationally, a topic detector based on a custom made dictionary or an existing one, such as Wikipedia, could detect such a situation where two different religions and religious beliefs are present, with no words that refer to historical or social events. Two possible approaches are specified below:

• The use of a dictionary to indicate certain topics and associated events within the paragraph. For example, the first dictionary may include religious topics marked by religious words. For example, Zorastrianism can have the following terms associated to it: zoroastrianism, ahura mazda, hadith, islam, sufism, madras, judaism, christianism, gospel, bible, torah, etc. A second dictionary can contain words that describe historical events such as: trade, war, siege, years, unification, evolved, etc. Lastly, another dictionary could contain historical entities such as: Alexander the Great, Cyrus, Persian Empire, etc. The dictionary could be built by hand, annotating key words automatically retrieved from older essays or created automatically using online resources They could be used to assess the sheer number of these topics in a paragraph and, based on analysis of previous essays, one can establish a ratio of these topics in paragraphs with specific characteristics.

• Alternatively, a machine learning approach can be used to detect a satisfactory combination of theology, history, anthropology, or social keywords used in a paragraph. The downside is that these methods require much hand annotated text.

One limitation of this approach is that the dictionaries and rules that need to be in place to detect the lack or abundance of contextual information, have to be tailored according to the domain of knowledge of the essay (or maybe even the specific topic of the essay).

3.2.3 Breadth of Background Research

Broad generalisations may be an indicator of narrow reading. Lists of entities and parts of speech (POS) tags around them can help determine the use of these entities or noun phrases. In practice, a program can look for lists of entities that are larger religions, countries and organisations; It can also look for modifiers, such as "all," next to nouns. Then, the program can check the frequency with which these entities "do" things or things that are "done" to them. If a general entity, such as Egypt or Islam actively affects ("do") other entities, such as "Sudan," a potential generalisation can be occurring in the text. Take for example the following two sentences:

(a) "*Egypt gave the Sudanese a chance for revolt and the very spontaneity of the Egyptian nationalists' revolt in Lower Egypt prevented their making any plans for the Sudan"*

(b) "*(...) those who had marginalised the Sufi master had not only failed to recognise the unifying role of Islam over the Sudan and its people.*"

In the first case (a), Egypt is *doing* something to the Sudanese people. It is giving it a chance of revolt. However, Egypt was divided at that time, so sentence (a) may be better phrased as "The political turmoil in Egypt gave the..." In the second example (b), Islam is attributed a unifying role in Sudan. That is, something *is done* to Islam (to attribute it a unifying role). In this case, the noun *Sufism* might have been a better, more precise choice of words. The student uses Sufism throughout the essay so he/she is probably looking for a different way of saying things. A warning might have forced him/her to choose a more precise word in this sentence.

3.2.4 Authoritative Sources for a Topic

Web searches can be used to provide an indicator of authoritative sources. For example:

• By checking references using Google's citation count or "link:" operator - more citations usually mean a better known source. For example: In one the essays analyzed, a student used two sources to explain the reasons that facilitate a Sudanese rebellion. The first source is Winston Churchill's book "The river war: an account of the reconquest of the Sudan" which is cited by 12 works in Google scholar. Contrast this with the second source cited: "The Mahdist state in the Sudan" by P.M. Holt which is cited by 119. Understandably, the grader noted that Churchill was not the most appropriate citation within this topic. Automatically, a citation count in google could have provided similar feedback.

• On a Google search, check the topic and the author of the reference. In the previous example, a web search of "churchill w. sudan" returns links that refer to Churchill fighting in Sudan whereas a search for "holt p.m. sudan" returns mostly links to works on his publications about the history of Sudan. An entity detection system such as OpenCalais[7] can provide meaningful insights on what can be computationally detected to establish differences between the the two queries. For example, OpenCalais reveals that the search using Holt P.M. returns a clearly detectable published work, whereas the search using Churchill does not.

We did note however that, in some subject areas, important material is not available in digital format, or only available via authenticated resources. In these cases care must be taken in using techniques that are based on online popularity.

3.2.5 Multiple Layers of meaning

For essays with complex events, background and implications, authors are prone to lose focus. The consequent layering of meanings and implications has been noted by the grader a few times in the sample essays. Some possible approaches to detecting this include:

• Counting the number of references by paragraph; too many may indicate lack of focus by citing too much and not making implications that narrow the claim or response.

• Counting the number of conjunctions such as "in addition," "moreover," etc. and the number of subordinating conjunctions such as "because" and "in order to/that." us.

These metrics would have to be fine tuned via analysis of a corpus of text comprised of similar essays to those being graded. Then, an appropriate ratio of these counts to each paragraph can be established.

3.2.6 Weakly Presented Arguments

To make a claim and support it, the wording should avoid ambiguities and weak or overly cautious interpretations.

• One approach to this is to count words such as "perhaps", "maybe", "possibly", "potentially", etc. and compare them with a preset threshold of frequency of such words. For example, the following paragraph from one of the essays would have benefited from a stronger wording in place of its overcautious "perhaps." Alternatively, the writer could have specified why she thinks "perhaps" is in order.

---

[7] http://www.opencalais.com/

"(...) draws heavily on Christian, Jewish and Zoroastrian influences, ***perhaps*** this again illustrates an intention to create a prophecy attractive to potential coverts."

This metric is widely used by other commercial software and should be appropriate for any kind of written essay.

3.2.7 Lack of Consistency Between Definitions and Use of a Word

Although this is very hard to do automatically, we think that concordances of keywords may enlighten the author. For example, if some word is accompanied by many adjectives which are not too similar ("good", "bad", "high", "low", etc.) then that word may be misused in some of those cases. Similarly, if the adjective is mostly accompanied by similar kinds of words but in one or two occasions it is not accompanied by one such word, it is possible that the word is being misused in that instance. Consider the following example.

• One essay contains the following three texts at different point in the essay: (a)"*(...)* ***Islam*** *over the* ***Sudan*** *and its people*;" (b) "*The* ***Mahdi*** *of the* ***Sudan*** *has to be(...);*" (c) "*(..) widely become known as the* ***Mahdi*** *of the* ***Sudan****, appears(...)*" By looking at words around Sudan, one could flag a potential confusion between *The Mahdi* and *Islam* because both are only two words apart from *Sudan*. This may lead the author to replace sentence (a) with something like " *influence of the Mahdi of the Sudan over his people*" if that is the intended idea.

3.2.8 Cognitive Choices Based on Word categories

Using dictionaries that allow categorisation of the text in multiple dimensions may be useful.

• LIWC categorises texts in over 70 dimensions. The authors of the software provide an online demo version[8]. This only includes a few categories, but it is easy to spot some potentially important differences - section 4.2 shows the results of a simple experiment on some sample essays.

3.3 Surface features of style

This section presents a sample of techniques which aim to warn writers about basic surface features of their writing that may result in confusing or poorly formatted text. No implications about writing intentions are drawn from these metrics.

Many of these techniques can be implemented simply by using a standard POS "tagger" followed by some simple post-processing - for example, LT-TTT2[9] provides natural language processing (NLP) components for a variety of text processing tasks such as tokenisation, sentence-splitting and rule-based entity recognition.

Dates (death, birth) when citing important people: It was observed that some students omitted references to a death date when referring to authors of relevant work As a quick metric; if the date of the writing is more than 80 years ago flag the author as a potential for a death date. Also, check Wikipedia for people's death date or compile a list of names in the field that must always be accompanied by a death date. Such names would probably be seminal writers, important historical figures, etc.

References in the correct format: Simple regular expressions can be used. For example, a loose detection of APA style citation can be achieved by using this regex: `([^\.].*?[0-9])(?=\.|\Z)`

Enough or too few references: Compute an average number of references per length of the essay and compare it to a threshold of the same metric computed from previous high quality essays.

Heavy use of nominals: Check the average ratio of entity names per paragraph and compare it to a threshold obtained from previous high quality essays

Stay within the word limit: Simple word count.

Outdated references: With regular expressions, a system can check reference dates and ask the user for a more recent reference if the year is earlier than a pre-set threshold year.

Too many prepositions: With a POS tagger, check the noun phrase (NP) to preposition ratio.

---

[8] http://www.liwc.net/liwcresearch07.php
[9] http://www.ltg.ed.ac.uk/software/lt- ttt2

Dangling modifiers: Using a POS tagger, check the number of NPs to pronoun ratio. An example would be: "*Sudan adopted Sufism because of its political structures.*" Here, the pronoun "*its*" can refer to Sudan or Sufism. In this case there are two nouns and one pronoun in the sentence. Without careful writing this becomes a recipe for confusion that could be avoided with an automatic warning.

Too much technical and complicated terminology: One could check noun phrases and "stopword" to word ratio and compared them to a previously computed threshold. In addition, the "long words" category in LIWC could be used in a similar way because long words is usually a proxy for complicated terminology.

Too many basic English words: Use the dictionary provided by Ogden's simplified English[10] to compute a ratio of simple words to length of the essay. Compare it then to a threshold that defines an acceptable use of simple words.

Check average length of sentences: Using a simple word count, check the median length of a sentence and the standard deviation of the sentence length. Again, this can be compared to a previously computed threshold.

Passive voice: Can be detected with a POS tagger. Again, determining a ratio of passives versus a pre-computed threshold can help a system assess the essay.

4 SOME EXAMPLES

This section presents two examples which illustrate the application of some of the techniques to real student text from the sample course.

4.1 TextTiling

TextTiling was used to compare the following two introductions separately:

• Title: Is the Bliss of the Beatific Vision in Paradise More Important in Sufism than the Fear of Hell?

The beatific vision of God in Paradise and the punishments of Hell each relate to God's dual aspects of mercy and wrath. Sufis have long been preoccupied with their souls' destiny, which relates to the question of God's essential nature: will He judge sins mercifully or will His wrath predominate? The Sufi master Ibn Al-Arabi asserted that God's nature and all of Islam can be conceived of in terms of the perspectives of 'tanzih' or God's transcendence, majesty and wrath, which is associated with rationality and 'tashbih' or God's immanence, beauty and mercy, related to the Sufi concepts of imagination or direct vision. Those following the Sufi path, which involves freeing the soul from its worldly attachments in order to be worthy of divine union through imaginative vision, would emphasise God's mercy over His wrath. This suggests that the bliss of the beatific vision in Paradise is more important in Sufism than the fear of Hell.

• Title: To what extent did Zoroastrianism, in particular the Ardā Virāf Nāmag, influence Islamic notions of paradise and hell?

Zoroastrianism was well established during the development of Judaism and well before the advent of Christianity and Islam. Located in Persia, at the cross-roads of civilizations, diffusion of religious ideas was common. The principal sacred text in Zoroastrian is the Avesta, among which are the Yasna (Y), the primary liturgical collection, thought to have been composed by Zoroaster himself. Although these text deals with Zoroastrian depictions of heaven and hell, it is the chapters of the Ardā Virāf Nāmag (AVN), a secondary work of semi-religious nature that developed them in detail. This narrates the story of a pious man, Viraf, who in a dream is taken on a journey to Heaven and Hell guided by two angles. Similarly the Qur'an has various depictions of paradise and hell including Muhammad's night journey; however, it is in Hadiths where they are elaborated in depth. These were later developed in more detail by scholars like Al-Ghazali (1058-1111). When evaluating the extent of Zoroastrian influence on Islamic notions of Paradise and Hell (Jahannam), the similarities and difference within these texts will be assessed.

The first text has a very good mark but the second has a very poor one. Moreover, the graders' comments explicitly say that the student fails to pose a motivating question.

[10] http://ogden.basic- english.org/

Analysing the text using TextTiling as detailed in Section 3.2.1, supports this assessment: the average score of inter-sentence similarity on the first essay is 0.18, much higher than the average score of the second essay: 0.09. Moreover, the similarity between the last sentence of the first essay and its title is 0.3 whereas the same similarity for the second text is only about 0.1.

4.2 LIWC

The online demo version of LIWC was used to compare the following two paragraphs; one from an poorly graded essay (54 points) and the other one from a very good one (80 points).

• *However the dualism of heaven and hell originates in the Yasnas which are generally dated to the second millennium B.C. Although, Islamic ideas of paradise and hell were also heavily influenced by the Jewish notions, Judaism itself was greatly influenced by Zoroastrianism. Form the above comparisons it is clear that directly or indirectly Zoroastrianism had a great influence on Islamic notions of heaven and hell. As the oldest of the revealed world religions Zoroastrianism, as Boyd describes, has 'had more influence on mankind, directly or indirectly, than any other single faith.'*

• *This explains why the Sufi master, Junayd, did not aspire to beatific vision because, to see God implies distance from Him and the true aim of Sufism is complete union with the beloved 13. Chittick argues that the annihilation of self is key to this union, as once the self is dead no unveiling of God is required because one will be able to see that everything in the universe (not just all of creation) is the immanent God Himself. This is what he calls the 'paradox of the veils'; the veils which obscure God actually make Him present 14. Thus I would argue that whilst the beatific vision is more significant in Sufism than the fear of Hell because God's mercy predominates, actually its key aspect is love of God for itself alone. Only when the Sufi loves God without any hope of future benefits such as the beatific vision, is the self completely annihilated and so, paradoxically, the beatific vision becomes possible in this life.*

In the above examples, the first essay contains no self references, a low occurrence of social words and a balance between negative emotion and positive emotion words. It also contains a very high incidence of long words. Contrast this with the second, better evaluated and more compelling essay. It contains a modest amount of self references, significantly more social words, a rather unbalanced emotional charge, slightly more cognitive words than the first essay and significantly less long words. This is surprising considering that it is noticeably longer than the first example.

5 CONCLUSIONS AND FUTURE DEVELOPMENTS

• By analysing comments and scores assigned to student writing (both wikis and traditional essays), we have developed some criteria and guidelines for producing good writing in the subject area of Divinity. These will be of help to students, when writing, and to markers in clarifying the criteria to be used. By automatically analysing the submitted texts, we have also identified a number of objective features which correlate with the "good" and "bad" writing.

• We have surveyed a range of computational linguistic techniques which have the potential to automatically evaluate the identified criteria. We conclude that this is practical in many cases, and the techniques often involve simple word counts, checks against dictionaries and checks against regular expressions.

• There is considerable potential to combine these techniques into a single tool which would provide valuable formative feedback to the students. Such a tool could be built to integrate independent modules which could be created and evolved by different developers, while presenting a consistent and integrated interface to the student.

REFERENCES

[1] P. Black and D. William. Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1):7 – 74, 1998.

[2] D. Boud. Assessment and learning: contradictory or complementary. In P. Knight, editor, Assessment for Learning in Higher Education, pages 35–48. London: Kogan Page/SEDA, 1995.

[3] P. T. Knight. Summative assessment in higher education: practices in disarray. Studies in Higher Education, 27(3):275 – 286, 2002.

[4] D. Carless. Learning-oriented assessment: conceptual bases and practical implications. In Innovations in Education and Teaching International, volume 44.1, pages 57–66. Routledge, 2007.

[5] G. Gibbs and C. Simpson. Conditions under which assessment supports students' learning. In Learning and Teaching in Higher Education, volume 1. 2004-5.

[6] D. J. Nicol and D. Macfarlane-Dick. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. Studies in Higher Education, 31(2):199 – 218, 2006.

[7] D. Hounsell. The trouble with feedback: New challenges, emerging strategies. In TLA Interchange. Centre for Teaching, Learning and Assessment, University of Edinburgh, 2008.

[8] P. Denton, J. Madden, M. Roberts, and P. Rowe. Students' response to traditional and computer-assisted formative feedback: A comparative case study. British Journal of Educational Technology, 39(3):486–500, 2008.

[9] S. Valenti, F. Neri, and A. Cucchiarelli. An overview of current research on automated essay grading. In Journal of Information Technology Education, volume 2, 2003.

[10] K. E. Kiefer and C. R. Smith. Textual analysis with computers: Tests of Bell Laboratories' computer software. Research in the Teaching of English, 17(3):201–214, 1983.

[11] D. E. Kieras. An advanced computerized aid for the writing of comprehensible technical documents. In B. Britton and S. Glynn, editors, Computer writing environments: Theory, Research, and Design. Erlbaum, Hillsdale, NJ, 1989.

[12] P. Wiemer-Hastings and A. C. Graesser. Select-a-kibitzer: A computer tool that gives meaningful feedback on student compositions. Interactive Learning Environments, 8(2):149–169, 2000.

[13] J. Robertson and P. M. Wiemer Hastings. Feedback on children's stories via multiple interface agents. In ITS '02: Proceedings of the 6th International Conference on Intelligent Tutoring Systems, pages 923–932, London, UK, 2002. Springer-Verlag.

[14] P. W. Foltz, S. Gilliam, and S. Kendall. Supporting content-based feedback in on-line writing evaluation with lsa. Interactive Learning Environments, 8(2):111 – 127, 2000.

[15] Y. Attali and J. Burstein. Automated essay scoring with e-rater v. 2. Journal of Technology, Learning, and Assessment, 4(3):1–31, 2006.

[16] T. Miller. Essay assessment with latent semantic analysis. Journal of Educational Computing Research, 29(4):495–512, 2003.

[17] M. A. K. Halliday and R. Hasan. Cohesion in English (English Language). Longman Pub Group, May 1976.

[18] J. L. Meurer. Relationship between cohesion and coherence in essays and narratives. Fragmentos, 25:147–154, July 2003.

[19] J. Bae. Cohesion and coherence in children's written english: Immersion and english-only classes. Applied Linguistics, 12(1):51–88, 2001.

[20] W. Dakka and L. Gravano. Efficient summarization-aware search for online news articles. In JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pages 63–72, New York, NY, USA, 2007. ACM.

[21] H. Halpin, J. D. Moore, and J. Robertson. Towards automated story analysis using participatory design. In SRMC '04: Proceedings of the 1st ACM workshop on Story representation, mechanism and context, pages 75–83, New York, NY, USA, 2004. ACM.

[22] J. W. Pennebaker and L. A. King. Linguistic styles: language use as an individual difference. Journal of personality and social psychology, 77(6):1296–1312, December 1999.

[23] A. C. Graesser, D. S. McNamara, and M. M. Louwerse. Methods of Automated Text Analysis. November 2010.

[24] M. A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. Comput. Linguist., 23(1):33–64, March 1997.

[25] S. Kaufmann. Cohesion and collocation: using context vectors in text segmentation. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pages 591–595, Morristown, NJ, USA, 1999. Association for Computational Linguistics.