Reflections on characteristics of university students' essays through experimentation with domain-independent natural language processing techniques

Debora Field debora@cs.ox.ac.uk University of Oxford
John Richardson john.t.e.richardson@open.ac.uk The Open University
Stephen Pulman stephen.pulman@cs.ox.ac.uk University of Oxford
Nicolas Van Labeke nicolas.vanlabeke@open.ac.uk The Open University
Denise Whitelock denise.whitelock@open.ac.uk The Open University

Abstract

This paper presents observations that were made about a corpus of 135 graded student essays by analysing them with a computer program that we are designing to provide automated formative feedback on draft essays. In order to provide individualised feedback to help students to improve their essays, the program carries out automatic essay structure recognition and uses domain-independent graph-based ranking techniques to derive extractive summaries. These procedures generate data concerning an essay's organisational structure and its discourse structure. We have selected 27 attributes from the data and used them in a comparative analysis of all the essays with a view to informing further development of the feedback program. The results of this analysis suggest that some characteristics of students' essays that our domain-independent feedback program is measuring may be related to the grades that tutors assign to their essays.

The SAFeSEA project (Supported Automated Feedback for Short Essay Answers) aims to develop an automated system to provide students with helpful and constructive feedback

on their draft essays. Educational research suggests that one particular type of feedback that falls within the scope of natural language processing – essay summarisation – is among the most useful for students (Nelson & Schunn, 2009). "Summarisation" includes both the traditional notion of a short précis and also simpler representations such as a list of an essay's key topics. As part of a larger prototype application, we have implemented essay structure recognition and key word and key sentence extraction procedures in a module that we call "EssayAnalyser". We have used the module to explore the attributes of a corpus of 135 essays that were produced by students taking a postgraduate course. This paper describes the results of that exploration and discusses the module's design.

Graph-based ranking methods

Our procedures are based on graph theory, which has been used in a wide variety of disciplinary contexts. The following account is based on that provided by Newman (2008). A graph consists of a set of *nodes* or *vertices* and a set of *links* or *edges* connecting them. (Some writers describe such a system as a *network*, but others restrict the latter to refer to graphs in which the edges are both directed and labelled.) A graph can be represented by an adjacency matrix in which the cells represent the connections between all pairs of nodes. In the simplest case, the cells take the value 1 if there is an edge between the relevant nodes and 0 otherwise.

Measures of *centrality* identify the most important or central nodes in a graph. They can therefore be used to measure how central (or key) a word, phrase, or sentence is in a natural language text of arbitrary length. The simplest such measure is *degree*, which is simply the number of edges attached to a node. Some other centrality measures take into account how strongly connected each node in the graph is to the whole graph, rather than just to its neighbouring nodes. We have used two of the latter centrality measures: eigenvector centrality (Brin & Page, 1998) and betweenness centrality (Freeman, 1977).

Text pre-processing and essay structure recognition

Before extracting key words and sentences from the text, the text is automatically pre-processed using modules from the Natural Language Processing Toolkit (Bird, Klein, & Loper, 2009). We also remove so-called "stop words" (articles, prepositions, auxiliary verbs, pronouns, etc.), which are the most frequently occurring in natural language but for our purposes the least interesting. We refer to the remaining meaning-rich words as "tidy" words and to the sentences without stop words as "tidied" sentences.

Structural components present in the essay are also automatically recognised and labelled (currently including preface, summary, abstract, introduction, discussion, conclusion, table of contents, quoted assignment question sentences, title, references, and appendices). This enables us to choose the sections of the essay that we wish to analyse for the presence of key sentences and key words. Currently only the prose of the body of the essay (introduction, discussion and conclusion) is considered.

As there are very few requirements in the assessment task and the assessment criteria concerning essay structure (only a word limit and referenced arguments and evidence from the literature), the essays in the corpus vary greatly in structure. They also vary in terms of text formatting choices that impact on structure. It was therefore decided that structure recognition would be best achieved without referring to a high-level formatting mark-up, and so the essays are converted to plain text files in UTF-8 encoding before they are processed by EssayAnalyser. The structure recognition rules have been hand-crafted from extensive experimentation with the corpus.

Key word extraction

Next, graph-based ranking methods are used to ascribe a "key-ness" rank to the lemma of each word in an essay. This follows Mihalcea and Tarau (2004), except that we use betweenness centrality to measure the centrality of a lemma in a text rather than eigenvector centrality. Key lemmas are defined as those in the top 20% of the ranked nodes that have betweenness centrality scores of .03 or more. (This threshold is where visual inspection identifies the sharpest bend in the "elbow" of the distribution curve in the key word centrality scores across all the essays.) The essay's key words are the inflections or base forms of the key lemmas that occur in the essay's original text. Key phrases are within-sentence sequences of key words that occur in the original text.

Key sentence extraction

Key sentences are also extracted using a graph-based ranking method. Instead of the lemmas, every true *sentence* in the essay is represented by a node in the graph. Each true sentence is then compared with every other true sentence, and a value is derived representing the semantic similarity of that pair. That similarity value becomes a weight that attaches to the edge that links the corresponding nodes in the key sentence graph. We are currently using cosine similarity as the similarity measure. The nodes are ranked using Mihalcea and Tarau's (2004) TextRank algorithm, and key sentences are defined as the top 30 ranked sentences. Note that no domain knowledge or other expert knowledge or "gold standard" model specific to a particular domain is used in the module's extraction of key words and key sentences.

Context

The essays were written by students taking a module entitled *Accessible online learning: Supporting disabled students*. This postgraduate module is presented annually over 20 weeks between September and January, it is worth 30 credit points (and thus equates to one quarter of a year's full-time study). The module is supported by a textbook (Seale, 2006) and by online resources, including links to a large number of external websites. Students are assigned to online tutorial groups and communicate with their tutors and one another by online forums. They are assessed by two assignments that are marked by their tutors and an end-of-module assignment that is marked by their tutor and an independent marker. All assignments are marked using a percentage scale on which the pass mark is 40%.

Assessment task

The first assignment is submitted online at the end of the first block, six weeks after the beginning of the module. The task requires that students discuss accessibility challenges for disabled learners in the student's own work context. Many students are professionals with extensive work experience in a wide variety of areas. This means that, although there is a set module textbook, student essays vary greatly in subject matter. A total of 135 students submitted the first assignment in 2010, 2011 and 2012. The EssayAnalyser program generated 27 characteristics of these 135 essays. These 27 attributes are listed in Table 1 together with brief explanations.

Table 1. Definitions of 27 attributes of students' essays (in alphabetical order)

Attribute name	**Definition**
% body == c	Percentage of the essay body (true sentences only) devoted to the conclusion section
% body == i	Percentage of the essay body (true sentences) devoted to the introduction section
all bigrams	Number of bigrams (made from key words)
all lemmas	Number of lemmas
all words	Number of words in the essay (occurring before the reference list or bibliography)
avfreq top5freq	Mean average frequency of the top five most frequent lemmas
avlen tidysent	Mean average length of a tidied sentence (a sentence without stop words in it)
bigrams in ass_q	Number of the essay's distinct bigrams that occur in the entire assignment question
c & toprank	Number of the top 30 key sentences that are in the conclusion section
distinct bigrams	Number of distinct bigrams
edges	Number of edges in the key sentence graph
edges/sents	Number of sentence graph edges divided by the number of true sentences
heads	Number of headings
i & toprank	Number of the top 30 key sentences that are in the introduction section
key lemmas	Number of key lemmas
key words	Number of key words
kls in ass_q_long	Number of essay's key lemmas occurring in whole assignment question
kls in ass_q_short	Number of essay's key lemmas occurring in assignment question's first sentence
kls in tb index	Number of essay's key lemmas occurring in module textbook index
len refs	Number of references in the references section
paras	Number of paragraphs
q sents	Number of sentences in whole assignment question quoted in the essay
sum freq kl_in_ass_q_long	Sum of the frequency counts (in the essay) for the essay's key lemmas that also occur in whole assignment question
sum freq kl_in_ass_q_short	Sum of the frequency counts for the essay's key lemmas that also occur in first sentence of assignment question
sum freq kls_in_tb_index	Sum of the frequency counts for the essay's key lemmas that also occur in the module textbook index
tidy words	Number of words in the essay ('all words') minus the stop words
true sents	Number of true sentences (excludes headings, captions, table of contents, title, etc.)

Exploratory factor analysis

An exploratory factor analysis was carried out on the values of these attributes for the 135 essays. A sample size of 135 is lower than the minima recommended by traditional texts (e.g. Comrey, 1973). However, more recent simulations have shown that robust results can be obtained from factor analyses with samples of 50–100 (Sapnas & Zeller, 2002) or even fewer (de Winter, Dodou, & Wieringa, 2009). First, a principal components analysis was used to determine the number of factors to extract. This identified nine components with eigenvalues greater than 1, and these explained 83.5% of the variance in the data. Nevertheless, the eigenvalues-greater-than-one rule is known to overestimate the true number of components in a data set because of sampling effects (Cliff, 1988). The bias is worse when the number of variables is large and the number of cases is small (both of which apply in the present case). Nowadays, it is generally acknowledged that the most reliable way to identify the number of factors in a data set is the parallel analysis of random correlation matrices. The analysis of 1,000 random correlation matrices was carried out using the program written by O'Connor (2000). The first seven components in the actual data set had eigenvalues greater than would be expected from a random data set, but the eighth and subsequent components did not. These seven components explained 74.7% of the variance in the data.

Principal axis factoring was therefore used to extract seven factors with squared multiple correlations as the initial estimates of community, and the extracted factor matrix was submitted to oblique rotation using a quartimin method. A cut-off of $\pm$.50 was used to identify those loadings that were salient for the purposes of interpretation. In Table 2, the variables with salient loadings on each factor are listed in descending order of the loadings in question. The resulting solution exemplified "simple structure" in that most of the variables loaded on one factor and only one variable loaded on more than one factor. The use of oblique rotation allowed for the possibility that the factors were correlated with one another. The correlation coefficient between Factor 2 and Factor 7 was .29. Otherwise, the correlation coefficients among the factors were all less than .20 in magnitude, implying that they were relatively orthogonal. It was therefore sensible to consider the variance explained by each factor.

Factor 1 explained 17.8% of the variance in the data set. Essays scored highly on this factor if (a) the frequency counts of the essay's top five most frequent lemmas were high compared to other essays; (b) the number of edges in the sentence graph relative to the number of true sentences was high; (c) there were relatively few key lemmas; (d) there were relatively few key words; (e) the number of edges in the sentence graph was high; and (f) there were relatively few key lemmas that also occurred in the module textbook index. This pattern would arise in essays with high average pair-wise sentence similarity but with low variation in word adjacency. We interpret this factor as reflecting the students' phrase structure creativity.

Factor 2 explained 13.4% of the variance in the data set. Essays scored highly on this factor if (a) the key lemmas in the short version of the assignment question occurred frequently in the essay compared to other essays; (b) the bigrams in the long version of the assignment question occurred frequently in the essay; (c) many key lemmas in the short version of the assignment question occurred in the essay; (d) the essay had many bigrams; and (e) the key lemmas in the long version of the assignment question occurred frequently in the essay. We interpret this factor as reflecting the students' attention to the terminology in the assignment question.

Factor 3 explained 9.5% of the variance in the data set. Essays scored highly on this factor if (a) there were many paragraphs; (b) there were many headings; (c) there were

Table 2. Loadings of 27 variables on seven factors (with salient loadings in bold)

Attribute	**1**	**2**	**3**	**4**	**5**	**6**	**7**
avfreq top5freq	**.89**	.06	.04	-.02	.01	-.03	.09
edges/sents	**.89**	.05	-.18	.04	.04	-.04	.18
key lemmas	**-.88**	.04	.03	.01	.00	.01	.08
key words	**-.85**	.09	.03	.10	.07	.01	.05
edges	**.74**	.04	.27	.36	.01	-.04	.11
kls in tb index	**-.64**	-.01	.10	-.02	-.08	-.01	**.61**
sum freq kls_in_ass_q_short	.24	**.95**	.06	.00	-.01	.10	-.14
bigrams in ass_q	.07	**.78**	-.01	-.08	-.01	.01	-.05
kls in ass_q_short	-.12	**.72**	-.04	.06	-.04	-.03	-.08
all bigrams	.02	**.59**	-.02	-.14	.01	-.02	.27
sum freq kls_in_ass_q_long	.46	**.50**	-.03	.05	.00	.00	.41
kls in ass_q_long	-.25	.44	-.06	.03	-.06	-.02	.36
distinct bigrams	-.25	.40	-.09	-.01	.07	-.03	.27
q sents	-.09	.30	.06	.14	-.01	-.06	-.07
paras	-.08	-.02	**.89**	-.01	.03	-.02	.17
heads	-.07	.02	**.72**	-.10	.05	.01	.13
true sents	.11	.01	**.70**	.40	-.05	-.01	-.07
avlen tidysent	-.02	.02	**-.56**	.08	.01	.06	.19
all lemmas	-.29	-.06	-.11	**.86**	-.04	.00	-.12
all words	.27	-.06	-.11	**.84**	.05	-.01	.08
tidy words	.19	.08	.21	**.80**	.00	.03	.19
len refs	-.10	.12	.06	.28	-.14	.02	-.01
% body == i	-.04	-.01	.02	.05	**.98**	.02	-.02
i & toprank	-.06	.06	.05	.03	**.96**	.00	-.03
c & toprank	-.05	.03	.01	.03	.04	**.90**	.01
% body == c	-.01	.00	-.01	.00	-.03	**.89**	-.01
sum freq kls_in_tb_index	.22	-.07	.07	.07	-.02	.00	**.88**

many true sentences; and (d) the tidied sentences tended to be short. We interpret this factor as reflecting the students' use of fundamental essay components. (Students who used more paragraphs and sentences would have to write shorter sentences to remain within the word limit.)

Factor 4 explained 10.7% of the variance in the data set. Essays scored highly on this factor if (a) the number of lemmas was relatively high; (b) the total number of words (including repeats) was relatively high; and (c) the number of tidy words (words after the

removal of stop words) was high. We interpret this factor as reflecting established properties of natural language (the average number of inflections per lemma occurring in English prose, and Zipf's law).

Factor 5 explained 7.6% of the variance in the data set. Essays scored highly on this factor if (a) a high proportion of the essay's true sentences occurred in the introduction; and (b) many of the top 30 key sentences occurred in the introduction. We interpret this factor as reflecting the quality of the introduction section.

Factor 6 explained 6.4% of the variance in the data set. Essays scored highly on this factor if (a) many of the top 30 key sentences occurred in the conclusion; and (b) a high proportion of the essay's true sentences occurred in the conclusion. We interpret this factor as reflecting the quality of the conclusion section.

Factor 7 explained 8.6% of the variance in the data set. Essays scored highly on this factor if the key lemmas in the textbook index occurred frequently in the essay; and (b) many of the key lemmas in the textbook index occurred in the essay. We interpret this factor as reflecting the students' attention to the terminology in the module textbook.

Finally, the regression method was used to estimate the scores obtained by the 135 essays on each of the seven factors. These factor scores are akin to standard scores (i.e. they have a mean of 0 and a standard deviation of approximately 1).

Analysis of covariance

An analysis of covariance was carried out to investigate whether these factor scores predicted the marks that the tutors had awarded the essays. The actual marks that were awarded to the 135 essays ranged from 24% to 88% with a mean of 63.7%. The analysis employed the students' gender as an independent variable and the seven factor scores as covariates. We noted that the number of references in the reference list had not shown a salient loading on any of the factors (see Table 2). Nevertheless, we considered that it might be an important predictor of the overall essay mark, and we therefore included it as an additional covariate.

The main effect of gender was not statistically significant, $F(1, 125) = 0.71$, $p = .40$, which indicated that there was no difference in the marks obtained by men and women when the effects of the covariates had been taken into account. There was a highly significant effect of the number of references, $B = .35$, $F(1, 125) = 11.09$, $p = .001$, which indicated that the students who cited more references tended to obtain higher marks. More specifically, for citing three extra references students would be expected to achieve an increase of 1 percentage point (i.e. $.35 \times 3$) in their overall mark. There was also a significant relationship with the scores on Factor 1, $B = 2.20$, $F(1, 125) = 4.78$, $p = .03$, which indicated that students who obtained higher scores on this factor also tended to obtain higher marks. Bearing in mind that most scores on this factor would fall within ± 2 standard deviations of the mean (i.e. between +2 and -2), the students with the highest scores would be expected to obtain marks 8.8 percentage points (i.e. $2.20 \times [2 - (-2)]$) higher than the students with the lowest scores. None of the other factor scores showed a significant relationship with the students' marks.

Conclusions

Our EssayAnalyser program uses state-of-the-art techniques in natural language processing to generate a rich description of the formal structure of students' essays without any domain-specific knowledge. The various attributes that it generates can be

reduced to a set of seven relatively independent constructs that explain a high proportion of the variance in the data set. Some of these constructs, especially Factors 4–7, can be explained by properties of mathematics, linguistics or program design. However, Factors 1–3 do not seem to be mere artefacts but reflect important aspects of how students go about writing essays, and one of these factors is a statistically significant predictor of the marks that their essays receive. We think they are worthy of serious consideration in future research.

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. Beijing: O'Reilly.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117.

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin*, *103*(2), 276–279.

Comrey, A. L. (1973). *A first course in factor analysis*. New York: Academic Press.

De Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, *44*(2), 147–181.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociology*, *40*(1), 35–41.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. In D. Lin & D. Wu (Eds.), *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404–411). Stroudsburg, PA: Association for Computational Linguistics.

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science*, *37*(4), 375–401.

Newman, M. E. J. (2008). Mathematics of networks. In S. N. Durlauf & L. E. Blume (Eds.), *The new Palgrave dictionary of economics* (2nd ed., Vol. 5, pp. 465–470). Houndmills: Palgrave Macmillan.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, and Computers*, *32*(3), 396–402.

Sapnas, K. G., & Zeller, R. A. (2002). Minimizing sample size when using exploratory factor analysis for measurement. *Journal of Nursing Measurement*, *10*(2), 135–154.

Seale, J. K. (2006). *E-learning and disability in higher education: Accessibility research and practice*. London: Routledge.