

Correlations between Automated Rhetorical Analysis and Tutors' Grades on Student Essays

Duygu Simsek¹, Ágnes Sándor², Simon Buckingham Shum³,
Rebecca Ferguson⁴, Anna De Liddo¹, Denise Whitelock⁴

¹Knowledge Media Institute
⁴Institute of Educational Technology
The Open University, Walton Hall,
Milton Keynes, MK7 6AA, UK

²Parsing & Semantics Group Xerox
Research Centre Europe 6 chemin
Maupertuis, F-38240 Meylan, France

³Connected Intelligence Centre
University of Technology, Sydney,
Broadway NSW 2007, Australia

{firstname.lastname}
@open.ac.uk

agnes.sandor
@xrce.xerox.com

Simon.BuckinghamShum
@uts.edu.au

ABSTRACT

When assessing student essays, educators look for the students' ability to present and pursue well-reasoned and strong arguments. Such scholarly argumentation is often articulated by rhetorical metadiscourse. Educators will be necessarily examining metadiscourse in students' writing as signals of the intellectual moves that make their reasoning visible. Therefore students and educators could benefit from available powerful automated textual analysis that is able to detect rhetorical metadiscourse. However, there is a need to validate such technologies in higher education contexts, since they were originally developed in non-educational applications. This paper describes an evaluation study of a particular language analysis tool, the Xerox Incremental Parser (XIP), on undergraduate social science student essays, using the mark awarded as a measure of the quality of the writing. As part of this exploration, the study presented in this paper seeks to assess the quality of the XIP through correlational studies and multiple regression analysis.

Categories and Subject Descriptors

K.3.1 [Computers and Education]: Computer Uses in Education

General Terms

Measurement, Design, Human Factors, Theory.

Keywords

learning analytics, writing analytics, academic writing, academic writing analytics, natural language processing, rhetorical parsing, metadiscourse, argumentation.

1. INTRODUCTION

In order to be regarded as academically literate, students in higher education need to engage with the ideas in the literature by

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org. LAK'15, March 16–20, 2015, Poughkeepsie, NY, USA. Copyright 2015 ACM 978-1-4503-3417-4/15/03...\$15.00 <http://dx.doi.org/10.1145/2723576.2723603>

recognising when significant claims are being made in articles, and by demonstrating the ability to examine them critically. One of the key requirements of academic writing in higher education is that students must develop a critical mind, make their thinking visible, and learn to construct sound arguments in their discipline [1].

Educators expect their students to learn to write in an academically sound way; specifically, learn to make knowledge-level moves and claims in their essays by recognising, and deploying, scholarly rhetoric, which is often articulated by *metadiscourse*. This term refers to the features of text that convey the author's intended meaning and intention [2]. Metadiscourse provides linguistic cues that explicitly organise the discourse, express a viewpoint, argument and claim, engage the readers, and signal the writer's stance [2]. For example, in Figure 1 the italicised words are the elements of metadiscourse used by the author to signal the rhetorical function of the sentences as summary.

The purpose of this article is to develop the idea that the procedures in any given classroom or laboratory exercise should be definitely determined by the specific aim, which the instructor has in mind to accomplish.

The perspective I shall use in this essay relies heavily on the view of professionalization presented in Andrew Abbott's brilliant study, *The System of Professions* (Abbott, 1988).

This paper explores social practices of propagating 'memes' (pronounced, 'meems') as a dimension of cultural production and transmission within Internet environments.

Figure 1. Metadiscourse that conveys summary statement

When assessing their students' writing therefore, educators will, among other features, be looking for scholarly metadiscourse as an indicator of argumentation. From a discourse-centric learning analytics perspective, a significant development in Natural Language Processing (NLP) is the automatic recognition of the rhetorical signals that authors use in research articles when making a significant scholarly move. Such powerful computational language technologies for extracting metadiscourse are becoming available; but since they are originally developed in non-educational contexts, there is a need to validate them in a higher education framework.

This paper describes the first study undertaken to assess the validity of a language technology, the Xerox Incremental Parser (XIP), on undergraduate social science student writing.

2. RELATED WORK

This section provides an overview of the relevant aspects of the three research areas related to our study: teaching academic writing, automated rhetorical analysis and learning analytics.

2.1 Teaching Academic Writing

The main purpose of an academic author is to convince readers of the validity of the claims put forward [3]. Consequently, as philosophers of science and learning science researchers have argued, rhetoric serves important functions within an argument, by both engaging readers with the claims that are being made and signaling their epistemic status [4].

However, educational research literature shows that students and educators have contrasting views regarding the expectations and interpretations of written assignments [5-8]. There is a general consensus among educators in all disciplines that the key element of student writing is “argument” [5, 8], whereas students think that the presentation of content and knowledge is the most important element, and that argumentation is one of the least important assessment factors for their educators [8, 9]. Research studies also show that high-scoring essays tend to be richer in argumentation, whereas low-scoring essays are more focused on factual descriptive information, and are poor in argumentation [8,10]. This indicates that educators value rhetoric and argumentation within student essays and look for their students’ ability to present and pursue well-reasoned, and strong arguments.

Textbooks and guides for scholarly style and argumentation are abundant; yet the acquisition of the skills remains a difficult task. New experimental methods could benefit from language technology tools that have been developed for the analysis of scholarly language.

2.2 Automated Rhetorical Analysis

There exist today some natural language processing systems that detect authors’ rhetorical moves in scholarly texts. One approach is argumentative zoning [11], which assigns a rhetorical move label to every sentence of the article. Another approach is concept matching [12], which assigns rhetorical move labels to rhetorically salient sentences only. Automated rhetorical analysis could be used both in evaluation, and as a self-teaching tool for students, who could inspect the rhetoric and argumentation of their own writing through the use of a rhetorical analyser.

In our experiment we used the concept matching framework, since it has the potential to focus students’ and educators’ attention to salient sentences conveying specific rhetorical moves related to argumentation around research problems. The framework is implemented as the rhetorical module of the Xerox Incremental Parser (XIP) [13]. It detects and labels rhetorically salient sentences in scholarly writing based on the identification of metadiscourse conveying the author’s rhetorical strategy. The labels are the following: SUMMARY (summarising the goals or results of the article), EMPHASIS (emphasising the importance of ideas), BACKGROUND (describing background knowledge necessary for understanding the article’s contribution), CONTRAST (describing tensions, contrasts between ideas, models or research directions), NOVELTY (conveying that an idea is new), TENDENCY (describing emerging research directions), and OPEN QUESTIONS (describing problems that have not been solved) [14].

One of the key innovations of the KSV was the use of flexible thresholds in the creation of network representations. This is what allowed us to create visualizations of LSA-based representations of texts. Rather than attempting to provide a two-dimensional layout based on the first few dimensions resulting from the matrix decomposition used in LSA, our approach has been to determine the similarities between documents based on the cosines between the vectors representing documents. A graph is then created in which the nodes correspond to the documents and the edges correspond to the LSA-based similarities between them. A force-directed layout then applied to the graph such that the positions of nodes represent a dimensionality representation minimize the distortion of the (very low dimensional) representation. This representation of a maximally connected graph typically lacks clarity, and in typical cases where there are tens or hundreds of nodes the graph is essentially unintelligible due to the large number of edges.

NOVELTY

CONTRAST

Figure 2. A sample XIP analysis

XIP is a candidate parser for evaluating and teaching academic metadiscourse, if it can be embedded in a more complete learning analytics approach.

2.3 Learning Analytics

Analysing written texts manually is a labour-intensive process, which is an increasing problem as massive-scale learning takes place online. Therefore, academic writing analytics research is burgeoning especially in the field of automated analysis of student writing [15].

Learning analytics offer the potential for automated, timely, and formative feedback. But although computational rhetorical parsing technology has been developed to analyse academic writing, it is barely deployed in educational contexts. Since this study will be the first step towards validating rhetorical parsing in an educational context, it is too early to propose employing this technology for summative assessment.

3. STUDY

We currently undertake a broad research project that seeks to explore the possibilities of applying the XIP rhetorical parser in an educational tool [16-18]. As part of this exploration this study seeks to investigate to what extent XIP is accurate and sufficient for detecting good academic writing in students’ essays given the teachers’ grade as an evaluation measure. We ask the following research questions: Is there a correlation between the salient sentences extracted by XIP and final grades? What are the rhetorical markers out of the salient sentences detected by XIP that are most promising as indicators of good academic writing in students’ essay? How accurate is the XIP output?

3.1 Dataset

The student writing used in this study was from one of the final year undergraduate education and arts modules of the UK based distance education university, The Open University (OU). In the EA300 Children’s Literature module, students studied key examples of novels, picture books, poems and creative performance produced for children aged between 3 and 18. Students read a selection of related critical material and consider major themes, issues and debates in the field. At the end of the module, students wrote 3000 word long essays for an assignment that required them to engage in depth with texts and approaches explored within the module.

1307 students submitted an essay, which were marked out of 100 by an Associate Lecturer (‘tutor’). All tutors used the same marking grid with guidelines provided by the module team. Students were assessed, in part, on their ability to think through the strengths and limitations of the materials they used, and to express this critical thinking clearly in their writing.

The marking grid prompted the tutors to consider six aspects two of which are especially in line with the metadiscourse XIP can identify: *approach to alternative explanations and arguments* and *construction of academic argument*. These aspects evaluate to what extent students are engaged critically with ideas coming from different sources, discuss alternative explanations, and produce well structured, coherent and persuasive arguments.

3.2 Methodology

XIP was used to analyse the 1307 student essays. The XIP analysis results were quantified by calculating the total number of salient sentences extracted by the parser and the numbers of each rhetorical sentence type. First, a correlational study was conducted with these analysis results based on the essays' marks. This was followed by a multiple regression analysis in order to understand the effect, if any, of each rhetorical sentence type on essay marks.

3.3 Correlational Study Results

The correlational study was conducted to understand whether there is any relation between the number of results from the XIP output and the marks of the essays.

A Pearson [19] product-moment correlation coefficient was computed to assess the relationship between the total number of salient sentences found by XIP in student essays and the mark of these essays. Correlation was computed as 0.190, which means a weak, positive correlation between the essay marks and the total number of salient sentences extracted by XIP. Thus increases in the total number of salient sentences are weakly correlated with increases in mark. Table 1 below shows the coefficient results for each rhetorical sentence type and the mark.

Table 1. Correlational Study Results for each Rhetorical sentence type

Rhetorical Sentence Type	Value of the Correlation Coefficient	Strength of the Correlation
CONTRAST	0.151	Weak
BACKGROUND	0.109	Weak
TENDENCY	0.025	No meaningful correlation
EMPHASIS	0.076	No meaningful correlation
NOVELTY	0.097	No meaningful correlation

Overall, no negative and no meaningful correlation was found with the sentence types TENDENCY, EMPHASIS, NOVELTY, SUMMARY and OPEN QUESTION. There was a weak, positive correlation between the essay mark and the total number of CONTRAST and BACKGROUND sentences.

Although these results give some insights into the correlations between XIP findings and marks, they do not tell the whole story. Besides the total number of salient sentences, the rhetorical type distribution is known; so that it can be used to interpret how strongly each sentence type affects the final mark. This was done through multiple regression analysis using all 1307 essays of the dataset.

3.4 Multiple Regression Analysis Results

In the multiple linear regression model, the mark of the essays was taken as the dependent variable and the number of salient sentences for each XIP category (TENDENCY, EMPHASIS, NOVELTY, SUMMARY, OPEN QUESTION, CONTRAST and BACKGROUND) marked up in the essays as independent variables.

The regression model proved to be highly significant. Following normal convention, $p \leq 0.05$ signifies a statistically significant result, and $p \leq 0.001$ is regarded as highly significant. The p value for this model was less than 0.001, which indicates that the model is statistically highly significant. It means that this is a strong evidence to further interpret how strongly independent variables help to explain the essay mark with the model.

Adjusted R^2 measures the proportions of the total variability in the dependent variable, which is explained by the independent variables of the model. For this model the adjusted R^2 was 0.048, which means that 4.8% of the total variability in mark was explained by the independent variables.

When each independent variable was analysed, we found that the two of the independent variables: CONTRAST and BACKGROUND are statistically highly significant and have explanatory power for the dependent variable essay mark (CONTRAST, $p \leq 0.001$; BACKGROUND, $p \leq 0.001$).

When unstandardized coefficients were examined for these two independent variables, the following interpretations were made:

- for a one unit increase in the number of **CONTRAST** sentences within essays, the model predicts that the dependent variable, essay mark, will increase between **0.498** and **1.078** points (calculated as $B \pm 2 * \text{Std.Error}$), holding all other independent variables fixed/constant.
- for a one unit increase in the number of **BACKGROUND** sentences within essays, the model predicts that the dependent variable, essay mark, will increase between **1.075** and **3.431** points, holding all other independent variables fixed/constant.

For the rest of the independent variables the p value was not significant, therefore they cannot be interpreted in the same way as CONTRAST and BACKGROUND.

We carried out an internal validation using a randomly selected subset of the overall data. We set IBM's SPSS statistical software package to randomly select half of the data, and ran the regression analysis on this. This produced exactly the same results. We are currently repeating the study with similar datasets to see whether we get the same results for external validation.

4. DISCUSSION

4.1 The performance of XIP on the student essays

The scope of this study did not allow us to carry out a large-scale evaluation of the performance of XIP. One of the authors has evaluated 225 automatically detected salient sentences, and found that 49 (22%) of them did not play the role of the scholarly argumentation in the essay. We could not measure the coverage, i.e. the percentage of actual salient sentences that are missed.

An important source of errors is related to the specificity of literary essays that the current version of XIP does not account for: Since these essays involve the analysis of literary work, the rhetorically salient expressions may also be parts of that analysis and not of the scholarly argumentation. The following sentence, which refers to the children's story, Peter Pan, illustrates such a non-rhetorical expression detected by XIP (underlined):

Wendy is not seen to challenge this role even when she is out of her comfort zone and enters Never land.

The performance of XIP could be improved by adapting the system to the domain in future work. However, the noise in the literary essays in our study does not amount to a proportion that would undermine the validity of the statistical correlations.

4.2 Relationship between the tutors' marking grid and the salient sentences

As we previously pointed out, the teachers' marking grid contains criteria for evaluating the essays according to six aspects, two of which are particularly in line with our framework: "Approach to alternative explanations" and "Construction of Academic argument". Thus it is probably these two aspects that underlie the correlations between the tutors' grades and XIP results on sentences labeled as CONTRAST and BACKGROUND. In this section we briefly describe the underlying relationship between the two aspects and the semantics of the two sentence types, and we attempt to show why the other sentence types have not shown significant correlation with the tutors' grades.

Sentences labelled CONTRAST capture the expression of tensions, contrasts between ideas, models or research directions, and the sentences labelled BACKGROUND make reference to relevant other work. Thus these two sentence types in XIP do indeed perform discourse functions that convey "alternative explanations", which in turn are organic parts of "academic argument" [20]. The following example illustrates this overlap:

Trites' analysis of young adult literature disputes Hunt's assertion by arguing that children's literature often affirms the child's sense of Self and personal power (Hewings, 2009, p.287).

The XIP analysis captures that there is tension within ideas by matching the expressions "analysis disputes" and "assertion by arguing", which suggest that the student is aware of alternative analyses of young adult literature. At the same time these expressions position the student's statement within the framework of her/his academic argumentation over approaches to the child's characteristics in children's literature.

The quantitative study did not show any statistically significant correlations between the marks and the other salient sentence types detected by XIP: SUMMARY, EMPHASIS, NOVELTY, TENDENCY and OPEN QUESTIONS. Taking into account the evaluation aspects and the object of the essays, we propose the following explanation.

The SUMMARY sentences merely convey the idea that the author summarizes her essay. Thus these sentences do not contribute to any of the evaluation aspects. Referring to new research directions, raising open questions, emphasising ideas as surprising, or important, and describing research tendencies are not usual discourse moves in literature analysis; these are elements of argumentation schemes in mainly empirical research.

While salient sentences do indicate the author's awareness of alternative analyses, and show efforts to develop scholarly argumentation, their mere presence does not imply that the alternative analyses are discussed at a sufficient level, or that the argumentation is sound, well-structured or coherent. It simply signals that the writer does convey some content on alternative analyses, and that this argumentation does treat the topic in a scholarly style. Still the fact that the number of salient sentences shows a correlation with the grades indicates that the more scholarly metadiscourse is present in a student essay the more it is likely that it gets a better grade in the evaluation.

4.3 Some Outliers

Whereas in the great majority of the essays the grade was correlated with the number of salient sentences detected by XIP, in some rare cases high grades were given by the tutors to essays with very few salient sentences, and conversely, low grades were given to essays with a relatively great number of salient sentences. A close look at some of these essays allows us to provide some insight into these cases.

We observed that the high graded essays with few salient sentences that we examined have a strikingly vivid and literary style, which does not strictly follow the patterns of concise scholarly communication. These essays convey a personal approach, show deep knowledge, and use unconventional expressions. Alternative explanations required by the marking grid are provided, however they are embedded into a particular narrative flow, in which the expression of contrast is distributed along several sentences (underlined). Consider the following extract:

As Hunt states 'sameness and difference is the essence of children's books; they have recurrent ideas' (2009a,p. 71). He goes on to cite ... [Here comes a list of examples.] But is this the only tradition the book breaks? Based upon the themes detailed above this essay will look at what similarities and differences A Monster Calls has to children's literature from the last 250 years, focusing particularly on Tom's Midnight Garden.

Instead of referring to the alternative arguments through expressions like as "contrasting analyses" or "critical debates", the author of this essay lays them out in several steps.

What we have observed in the case of low-graded essays containing a relatively high number of salient sentences is that on the contrary, their style is simple and schematic, and sometimes their syntactic structure is not clear:

I do not think any of the themes I have mentioned were written about to change or challenge aspects of the community, I believe these issues were just to define the culture of society as it was in the Victorian era and to reinforce the roles subliminally.

In order to discover these outliers, and categorize them correctly, we would need to supplement rhetorical analysis with features that take into account style.

5. CONCLUSION

This study is an example learning analytics approach that can be followed by the wider LA community who might want to evaluate the potential use of analytics products within learning contexts,

for which there is a growing interest. The purpose of our study has been the evaluation of the rhetorical parser, XIP, in correlation with tutors' essay grades as a measure of quality.

We have made a significant advance toward understanding the power and effectiveness of XIP in educational contexts. The results show that the output of XIP is strongly related to teachers' expectations in student essays: we have found statistically significant correlations between two of the XIP rhetorical move labels, CONTRAST and BACKGROUND, and the final grade of the essays. These two labels convey rhetorical moves that are particularly in line with two aspects of the tutors' marking grid.

Following the quantitative study, we had a closer look at some student essays, which provided some insights regarding the parser's performance. We have found that the quality of the parser is reasonably good, considering that it has not been customized for the particular domain. We have also analyzed some essays, which got high grades and contain few salient sentences, as well as ones that got low grades and contain many salient sentences. Such essays show special writing style, and further work is needed to recognise them, and integrate their features into an automatic analysis tool.

Overall, the focus of this research is not on grading student writing automatically, but on the potential to automatically identify attributes of good academic writing, so that we can design computer-aided support for both educators and students in monitoring students' progress and in displaying the rhetorical analysis of the essays as formative feedback. Social science student essays have constituted the material of our first study, which we plan to follow up with student essays from various other disciplines. This will then allow us to create a framework that will be the 'middle ground' between learning and computation, helping members of both communities articulate, in precise terms, the opportunities for pedagogically sound learning analytics.

6. REFERENCES

- [1] S. M. Glynn and K. D. Muth, "Reading and writing to learn science: Achieving scientific literacy," *Journal of Research in Science Teaching*, vol. 31, pp. 1057-1073, 1994.
- [2] K. Hyland, *Metadiscourse: Exploring interaction in writing*: Continuum International Publishing, 2005.
- [3] A. de Waard, S. Buckingham Shum, A. Carusi, J. Park, M. Samwald, and Á. Sándor, "Hypotheses, evidence and relationships: The HypER approach for representing scientific knowledge claims," 2009.
- [4] W. A. Sandoval and K. A. Millwood, "The quality of students' use of evidence in written scientific explanations," *Cognition and Instruction*, vol. 23, pp. 23-55, 2005.
- [5] M. Lea and B. V. Street, "Student writing in higher education: An academic literacies approach," *Studies in higher education*, vol. 23, pp. 157-172, 1998.
- [6] R. Andrews, *Argumentation in Higher Education: Improving practice through theory and research*: Routledge, 2010.
- [7] T. Lillis and J. Turner, "Student writing in higher education: contemporary confusion, traditional concerns," *Teaching in Higher Education*, vol. 6, pp. 57-68, 2001.
- [8] L. S. Norton, "Essay-writing: what really counts?," *Higher Education*, vol. 20, pp. 411-442, 1999.
- [9] D. Hounsell, "Essay planning and essay writing," *Higher Education Research and Development*, vol. 3, pp. 13-31, 1984.
- [10] C. Coffin, M. J. Curry, S. Goodman, A. Hewings, T. Lillis, and J. Swann, *Teaching academic writing: A toolkit for higher education*: Routledge, 2002.
- [11] Teufel, S., Kan, M-J. (2009). Robust argumentative zoning for sensemaking in scholarly documents. In Proceedings of the 2009 international conference on Advanced language technologies for digital libraries (NLP4DL'09/AT4DL'09).
- [12] Sándor, Á. (2007). Using the author's comments for knowledge discovery. Semaine de la Connaissance: Atelier Texte et Connaissance. Nantes. June 29.<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.73.1938&rep=rep1&type=pdf>
- [13] S. Ait-Mokhtar, J.-P. Chanod, and C. Roux, "Robustness beyond shallowness: incremental deep parsing," *Natural Language Engineering*, vol. 8, pp. 121-144, 2002.
- [14] Sándor, Á. and Vorndran, A. (2010). The detection of salient messages from social science research papers and its application in document search. Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology, Buenos Aires, Argentina, May 10-14. 2010.
- [15] White, B., & Larusson, J. A. (2015). Chapter 8: Identifying Points for Pedagogical Intervention Based on Student Writing: Two Case Studies for the "Points of Originality". In J. A. Larusson & B. White (Eds.), *Learning Analytics: From Research to Practice* (156-190). Springer.
- [16] Simsek, D., Buckingham Shum, S., De Liddo, A., Ferguson, R. and Sándor, Á. (2014) Visual Analytics of Academic Writing, Demo at The 4th International Learning Analytics and Knowledge, Indianapolis, IN, USA, pp. 265-266, ACM New York, NY, USA ©2014
- [17] Simsek, D., Buckingham Shum, S., Sándor, Á., De Liddo, A. and Ferguson, R. (2013). XIP Dashboard: Visual Analytics from Automated Rhetorical Parsing of Scientific Metadiscourse. 1st International Workshop on Discourse-Centric Learning Analytics. (3rd International Conference on Learning Analytics & Knowledge, 8 April 2013, Leuven, Belgium). Open Access Eprint: (<http://oro.open.ac.uk/37391>)
- [18] Taibi, D., Sándor, Á., Simsek, D., Buckingham Shum, S., De Liddo, A. and Ferguson, R. (2013) Visualizing the LAK/EDM Literature Using Combined Concept and Rhetorical Sentence Extraction, 1st Learning Analytics and Knowledge Data Challenge at Learning Analytics and Knowledge (LAK '13), Leuven, Belgium
- [19] Pearson, K. (1895), *Royal Society Proceedings*, 58, 241
- [20] Swales, J.M., Feak, C. (1994). *Academic Writing for Graduate Students*. Ann Arbor, the University of Michigan Press Suthers D. and Verbert, K. Learning analytics as a "middle space". Proc. 3rd Int. Conf. on Learning Analytics and Knowledge (LAK '13), Leuven, BE, pp.1-4, ACM: New York, 2013.