Analysing student responses: early lessons from a pilot study

Jenny McDonald
University of Auckland
j.mcdonald@auckland.ac.nz

Adon Christian Michael Moskal
Otago Polytechnic
adon.moskal@op.ac.nz

Irina Elgort
Victoria University of Wellington
irina.elgort@vuw.ac.nz

Cathy Gunn
University of Auckland
ca.gunn@auckland.ac.nz

ABSTRACT: We report on early lessons from a pilot study to evaluate a new web-based tool for teachers. The tool is designed to support the rapid analysis of written student responses to short answer questions and was conceived to support formative assessment, especially in large-class settings, as well as to provide insights into teaching and learning design. We describe our approach to building the tool through working in partnership with academic developers and teaching staff from diverse academic contexts. We then discuss the challenges and opportunities that this process has presented. Insights from the pilot study to date suggest that the affordances of existing NLP technologies can be deployed to the advantage of teachers, and ultimately learners, by making the analysis of student responses in a range of contexts easier, quicker, more robust and transparent.

Keywords: Formative assessment, text analytics, automated short answer question marking, educational technology development.

1 INTRODUCTION

Beyond summative assessment of written student responses to questions, the rapid analysis of student responses has potential not only to provide opportunities for formative assessment but also to help teachers to assess when and in what ways teaching and student learning can be enhanced (McDonald, Bird, Zouaq & Moskal, 2017). Text is arguably central to teaching and learning; from deep questions designed to reveal how students understand and describe the world, to more surface-level questions relating to content knowledge. In addition, students' communicative ability is often assessed through examination of syntax and style. Yet, particularly in large class settings, it is challenging for educators to know what and how their students are learning. The time and resource constraints of modern tertiary environments mean it is often only at the time of marking final examinations that student understanding becomes clear to the teacher, by which time it is too late to respond to misconceptions.

1.1 Background

Automated assessment of student responses is one approach to addressing these constraints; to achieve outputs which are on a par with human graders is both an active research area and a work in progress (Burrows, Gurevych, & Stein, 2015). However, the goal of fully automated assessment can shift depending on how the assessment is conceived and structured. For example, assessing short responses to questions designed to check recall of facts is an easier task for both humans and machines than assessing longer responses to deeper or open-ended questions (Dzikovska, Nielsen & Leacock, 2015). Formative assessment involves more than assigning a single label or grade to a student response. In practice, multiple labels may be required to adequately categorise a single response (McDonald, Bird, Zouaq & Moskal, 2017). Reliability, both between human graders and within a single grader is also an issue (Elton & Johnston, 2002; Jonsson & Svingby, 2007)—if humans struggle to ascribe a consistent meaning to a response, even with the aid of scoring guides and rubrics, it is hard to see how an automated system will fare better. Further, in higher education settings, teachers who take an interpretivist approach may contest the notion of reliability itself (Orr, 2007). Finally, good teaching practice dictates that assessments can and should change regularly which adds a further layer of complexity for automated assessment, where supervised methods are used, since training data will be limited.

To address some of these issues, recent studies have adopted a more nuanced approach. Automatic methods are used to support human decision-making rather than replace it (e.g. Basu, Jacobs & Vanderwende, 2013). This provides several advantages. There is potential to improve marker consistency and objectivity as well as portability between contexts (Pado & Kiefer, 2015). Teaching insight can be gained and pedagogenic errors—errors induced by the teacher or by the teaching environment (Laurillard, 2002)—identified by looking for co-occurrence of text fragments between responses and between responses and teaching materials (McDonald, Bird, Zouaq & Moskal, 2017). A key advantage is that the analytic process, and indeed the learning process, represented by student text is made more transparent for the teacher. This not only helps to assuage concerns about 'black-box' algorithms replacing human input (in particular where high-stakes assessment is concerned, e.g. Ericsson & Haswell, 2006; Pado & Kiefer, 2015), it also provides a window into student understanding, and thus the opportunity for a *teaching moment* (Havighurst, 1952).

1.2 A text analysis tool for teachers

Consistent with the use of text analytic methods to support human assessment we describe a pilot study to evaluate a new text analysis tool for teachers. Quantext is an online platform designed to help teachers to extract insights from student responses to short-answer questions (McDonald & Moskal, 2017). While the following description outlines the key features of Quantext in terms familiar to natural language processing (NLP) researchers and practitioners, it is important to note that the Quantext interface has been designed for the non-specialist. The use of jargon is avoided, and where it is used, is explained through in-context tooltip displays.

Teachers upload student responses to Quantext, which extracts text features from the dataset, and aggregates and presents them in a variety of forms for further exploration. For example, Quantext first displays the most frequent words, bigrams (two word units) and trigrams (three word units) from the student responses. Teachers can then click on a word or multi-word unit of interest, and explore how it is being used by students via a keyword-in-context display and associated wordtree visualisation (Wattenberg & Viégas, 2008). Quantext allows teachers to label responses via a sorting process based on selected features such as response length, words or ngrams (multi-word units), as well as on readability indices and semantic similarity. Labels are created by teachers and multiple labels can be applied to any given response. Text-based teaching materials, such as lecture transcripts, handouts or course books can be uploaded as a reference corpus and features common to both student and teacher discourse can be highlighted. Stopwords, choice of common readability indices, algorithms for calculating ngram keyness, and semantic similarity measures are all in control of the teacher. Finally, Quantext has been designed specifically to support comparison of analyses both within and between student cohorts.

In the next section, we describe our approach to building the tool through working in partnership with academic developers and teaching staff from diverse academic contexts. Specific design decisions arising from this partnership are highlighted.

2 APPROACH TO QUANTEXT DEVELOPMENT – A PILOT STUDY

2.1 Starting from student responses

The need for a tool like Quantext became apparent through a series of New Zealand-wide workshops designed to introduce existing text analysis tools to teachers. The workshops were part of a NZ-wide Ako Aotearoa funded project, *Building an evidence-base for teaching and learning design using learning analytics* (Gunn & McDonald, Forthcoming). Teachers and learning designers who attended the workshops expressed enthusiasm for analysing text and a willingness to try existing and readily available tools (e.g. http://www.laurenceanthony.net/software/antconc/ (Anthony, 2014), http://www.sketchengine.co.uk (Kilgariff et al., 2014), and Textalyser.net). However, while in principle the idea was well-received, in practice it was clear that existing text analytic tools presented numerous obstacles for teachers unfamiliar with concepts spanning linguistics, computing, data management and statistics. From feedback during the workshops, it was clear that the design and purpose of existing tools was not well-aligned with the specific needs of most teachers. This is not a criticism of the tools themselves, far from it; rather, it is merely a reflection of the fact that they were not designed to meet the specific needs of teachers seeking insights from student generated text. We therefore directed our efforts towards learning from the experience and creating a tool geared towards the average teaching academic. Informed by the findings from an exploratory study situated in an undergraduate health sciences context (McDonald, Bird, Zouaq & Moskal, 2017), we created a proof of concept in the form of a Jupyter notebook. While still far from accessible to most teachers, the basic functions of the notebook involved iterative sorting of student responses based on simple text features such as response length and key words and visualising the results. This approach resonated with academic developer colleagues and from there the first iteration of Quantext as a web-based tool developed.

2.1 Working in partnership with teaching staff

While still in a rudimentary stage, and in partnership with academic developers who work within tertiary institutions to support teacher professional development, we recruited a small number of teachers to a pilot study of the fledgling tool. The 8 pilot participants so far have been drawn from 4

NZ tertiary institutions; 3 universities and 1 polytechnic. Both the teachers, and their academic developer guides, provided early evaluation and input into the iterative development of the tool. Tertiary teachers involved in the pilot are from diverse disciplines with the majority interested in analysing text from their undergraduate classes. The range of disciplines includes Physics, Philosophy, Architecture and Design, and Medicine with class sizes ranging from around 100 students to more than 2,000. In addition, some pilot participants were interested in trying out Quantext with data from Massive Open Online Courses (MOOCs)—one is a statistics MOOC offered through the University of Auckland (approximately 20,000 students at the start of the course), while the other is a MOOC on Antarctica offered through Victoria University of Wellington (approx. 2,000 students at the start of the course – Elgort, Lundqvist, McDonald & Moskal, 2018). Students may join or leave the MOOCs at any time.

A key feature of the pilot is that while all teachers came to it with an interest in analysing student text, they also came with questions, assignments, student responses from earlier cohorts, and in some cases discussion forum posts; data specific to their context and not designed to test the tool. In other words, the context came first rather than tool development. In this way we planned to iteratively refine the tool in order to address specific teaching needs. At the time of writing the pilot study is still underway and planned to continue through semester 1, 2018. The broad goals of the pilot study are to evaluate the utility of Quantext along several dimensions (e.g. speed of analysis, validity and reliability of data and so on) and report on teacher reflections from tool use. While still at an early stage, we have already identified challenges and opportunities revealed by the pilot study to date. We describe these challenges and opportunities in the final sections of this paper.

3 CHALLENGES

3.1 Data input

We knew from interviews with teachers, and survey results conducted as part of the Ako Aotearoa project, that getting data into and out of systems often presents a challenge for teachers. For the purpose of the pilot study we therefore standardised the data input format to a Microsoft Excel spreadsheet—questions are listed in the first worksheet of the spreadsheet, one question per row; student ids and responses are listed in subsequent worksheets which are numbered according to their corresponding question number. While in the short-term this simple approach has resulted in some calls for help, in the main any issues have been quickly resolved and have the benefit that once uploaded, all data is in a consistent format. Eventually, input of question and response data should be directly integrated with Learning Management Systems (LMS) and other common assessment platforms.

3.2 Question length and style

An associated challenge was handling the range in style of question and lengths of responses. Question style so far has ranged from single questions to multiple part and sub-part questions. The ability to retain context across related questions is important and needs to be handled. In our earlier work, student responses had been less than 50 words in length with an average response length of around ten words. By contrast, the average response length from early pilot data was closer to 200 words. While this was not an issue in terms of data format for uploading, or for our processing engines, it did prove an issue for how to display responses in the fledgling Quantext interface. In particular, our spreadsheet view became unwieldy, as these longer responses resulted in excessive scrolling down the page to view all responses. These issues are currently being addressed.

3.3 Specialist terminology

Although we took care to reduce specialist terminology wherever possible in the Quantext interface, even the most basic terms in everyday use can become problematic when they also have a specific technical meaning. For example, in corpus linguistics, a 'keyword' is one which occurs more often than expected by chance and is calculated by comparing word frequency between a given text and a reference corpus. In common use, 'keyword' simply means a word which represents the central meaning of a text. It became apparent from discussions early in the pilot that these differences need to be made explicit, or alternative terms chosen, in order to avoid confusion. Key phrases and blacklist words are other examples.

3.4 Accessible student response corpora

We found one of the most useful things, when recruiting participants to the pilot study, is to demonstrate the tool with an authentic dataset and one which resonates with participants. Furthermore, such datasets, in particular ones which include teacher annotations or categorization, are invaluable to evaluate both interface usability and unsupervised metrics such as similarity scoring. However, there are few publicly accessible datasets available for this purpose. One example is of student responses to questions in an undergraduate computer science course (Mohler & Mihalcea, 2009), and a second contains questions relating to understanding of US civics (Basu, Jacobs & Vanderwende, 2013). There are also some limited datasets available on Kaggle (https://www.kaggle.com/datasets). In practice, we have found far greater variability in terms of question format, number of responses and response length from the data we have observed in the pilot study thus far than from existing publicly available datasets. We are therefore seeking permission to release (anonymised) datasets obtained through this project to the wider research community.

3.5 Cost of development and support

Development and support costs can present a challenge. To some extent we have overcome these issues through developing Quantext as a sideline to our day-jobs. We hope that by releasing Quantext as an open source project this will attract wider interest and additional resource. In the meantime, alignment of pilot study goals with individual teacher/academic developer research interests has contributed to keeping costs down, however additional resource will be required to substantially progress the project.

Time and workload are further costs. Tertiary teachers have many demands made of them daily. Contributing to any pilot study takes time and of necessity is fitted in around other work. Learning to use a new tool and contribute feedback on it may result in tasks taking longer than they otherwise would. This can be counterproductive; in the early stages, a tool designed to save time may paradoxically take more time to use. For this reason, among others, pilot studies such as this one inevitably tend to recruit and indeed rely on, highly motivated teachers who may not necessarily represent the wider user cohort. This can have an impact on wider adoption (e.g. Gunn, Woodgate & O'Grady, 2005).

3.6 Integration with institutional systems and LMS

Consistent feedback from pilot study participants has pointed to the desirability of developing plugins to use Quantext within institutional LMS or leveraging LMS APIs to automatically populate Quantext. The ideal is for teachers to set questions within an LMS or other assessment tool and have student responses automatically available in Quantext. Following analysis it makes sense to export categorised output directly to tools such as the Student Relationship Engagement System - SRES (Liu, Bartimote-Aufflick, Pardo & Bridgeman, 2017) which would facilitate automating feedback, based on assigned labels, directly to students. LMS integration will almost certainly reduce the length of time taken for analysis. A goal of the pilot is to evaluate the speed of analysis independently of LMS integration. Assuming a positive evaluation, we plan to implement integration enhancements at the conclusion of the pilot project.

4 OPPORTUNITIES

4.1 Professional teacher development

The importance of academic developers working with teachers to explore the use of Quantext (or arguably any learning analytic tool) in its early stages of development, cannot be overstated. This assists with ensuring teacher and learner needs are appropriately addressed and the potential for teaching improvement and enhancing formative assessment is realised. There are opportunities to help teachers with analysing textual data in general and with writing effective short answer questions. For example, ambiguous questions can be quickly identified when the most frequent words or multi-word units in responses turn out to be completely unexpected. Furthermore, comparison of student responses with teaching materials allows teachers to see directly the impact of some of their teaching choices and learning designs reflected in student responses. This in turn encourages reflection and action (Schön, 1987).

4.2 Analytics for online fora and student evaluations

As a result of requests for functionality arising from the pilot we need to decide whether to extend Quantext to handle other forms of text-based data common in a higher education context. For example, currently Quantext is ill-equipped to handle online discussion forum data; Quantext treats student responses as independent, but forum data is highly dependent on the relationships between posts (e.g. post order, or whether a post is a direct reply to another). Incorporating such data into Quantext, other than by treating posts as independent responses (Elgort et al., 2018), will necessitate changes to the way in which data is stored and referenced, and changes to the interface to display the relationships inherent in the data. Another common data source in teaching and learning contexts are the free text comments sections of student evaluations and surveys. There is also scope to explore the use of Quantext as a research tool for analyzing qualitative elements of surveys and interview data. While Quantext can currently process this data and present teachers with summary statistics, additional NLP techniques such as sentiment analysis may be useful additions to the featureset.

5 CONCLUSION

End-user text analytic tools and platforms must be able to deal flexibly with data sources, be robust, easy to use and fast (Ittoo, Nguyen, & van den Bosch, 2016). Quantext for academic contexts currently aligns well with these aims. In addition, the analysis of student text can and should benefit from the input of teachers. Early indications from this pilot study suggest that there is no need to wait for NLP approaches to fully automated processing to achieve comparable accuracy to human markers, if indeed this is desirable in practice and can be achieved. The affordances of current advances in NLP technologies can be deployed to the advantage of teachers and ultimately learners by making the analysis of student responses in a range of contexts easier, quicker, more robust and transparent.

REFERENCES

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: A clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics,* 1, (pp. 391–402).

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education, 25*(1), (pp. 60-117).

Dzikovska, M. O., Nielsen, R. D., & Leacock, C. (2015). The joint student response analysis and recognizing textual entailment challenge: Making sense of student responses in educational applications. *Language Resources and Evaluation*, 1–27.

Elgort, I., Lundqvist, K., McDonald, J., & Moskal, A.C.M. (2018). Analysis of student discussion posts in a MOOC: Proof of concept. *Companion Proceedings 8th International Conference on Learning Analytics & Knowledge (LAK18).*

Ericsson, P. F., & Haswell, R. H. (2006). *Machine scoring of student essays: Truth and consequences*. Utah State University Press.

Elton, L & Johnston, B (2002) *Assessment in Universities, a critical review of research*. LTSN Generic Centre. Available online at https://eprints.soton.ac.uk/59244/1/59244.pdf

Gunn, C., & McDonald, J. (in press). Promoting learning analytics for tertiary teachers: A New Zealand case study, In J. Lodge, L. Corrin & J. Cooney Horvath (Eds.), *Learning analytics in the classroom: Translating research for teachers.* Routledge.

Gunn, C., Woodgate, S., & O'Grady, W. (2005). Repurposing Learning Objects: A Sustainable Alternative? *Association of Learning Technology Journal, ALT-J, 13*(3), (pp. 189-200).

Havighurst, R. J. (1953). *Human development and education*. Oxford, England: Longmans, Green.

Ittoo, A., Nguyen, L.M., & van den Bosch, A. (2016). Text analytics in industry: challenges, desiderata and trends. *Computers in Industry, 78*., (pp. 96-107).

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, *2*(2), (pp. 130-144).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovvář, V., Michelfeit, J., Rychlý, P & Suchomel, V. (2014) The Sketch Engine: ten years on. *Lexicography*, 1. (pp. 7-36)

Laurillard, D. (2002) *Rethinking University Teaching: a framework for the effective use of educational technology*. RoutledgeFalmer, London.

Liu, D.Y.-T., Bartimote-Aufflick, K., Pardo, A. & Bridgeman, A.J. (2017). Data-Driven Personalization of Student Learning Support in Higher Education. In Peña-Ayala, A. (Ed.) *Learning Analytics: Fundaments, Applications, and Trends: A View of the Current State of the Art to Enhance e-Learning*. Springer. (pp. 143–169)

McDonald, J., Bird, R.J., Zouaq, A. & Moskal, A.C.M. (2017) Short answers to deep questions: supporting teachers in large-class settings. *Journal of Computer Assisted Learning* 33(4), (pp. 306–319). doi: 10.1111/jcal.12178

McDonald, J. & Moskal A.C.M. (2017). Quantext: analysing student responses to short-answer questions. In H. Partridge, K. Davis, & J. Thomas. (Eds.), *Me, Us, IT! Proceedings ASCILITE2017: 34th International Conference on Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education* (pp. 133-137).

Mohler, M., & Mihalcea, R. (2009*).* Text-to-text semantic similarity for automatic short answer grading*. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 567-575). Association for Computational Linguistics.

Orr, S (2007) Assessment moderation: constructing the marks and constructing the students, *Assessment & Evaluation in Higher Education*, 32(6), (pp. 645-656), DOI: 10.1080/02602930601117068

Pado, U. & Kiefer, C. (2015). Short answer grading: When sorting helps and when it doesn't. *Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning at NODALIDA 2015*. NEALT Proceedings Series 26 / Linköping Electronic Conference Proceedings 114 (pp. 42–50).

Schön, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. Jossey-Bass.

Wattenberg, M. & Viégas, F.B., (2008). The word tree, an interactive visual concordance. *IEEE transactions on visualization and computer graphics*, *14*(6)