Automatic Generation of Student Report Cards

Amy Isard
School of Informatics
University of Edinburgh
Scotland, U.K.
amy.isard@ed.ac.uk

Jeremy Knox
School of Education
University of Edinburgh
Scotland, U.K.
jeremy.knox@ed.ac.uk

Abstract

The Learning Analytics Report Card (LARC) is a pilot system which takes time-series data from a student's course-related activity in a Virtual Learning Environment and generates automatic textual summaries in real time. Students are able to generate reports as often as they like, and to choose which aspects of their behaviour are included in each report. As well as rating a student's scores against set standards, the generated texts make comparisons with the individual student's previous behaviour from the same course, and with the average scores of their student cohort. In addition, we carry out sentiment analysis on the student's forum posts, and generate a summary using quantifiers. We report some student reactions to initial trials of the system.

1 Introduction

The Learning Analytics Report Card (LARC) project was an interdisciplinary pilot project at the University of Edinburgh involving researchers in Education and Computational Linguistics, as well as Information Services. Its overall aim was to raise students' critical awareness of the ways in which learning analytics (Ferguson, 2012) can intervene in and mediate educational activity. The project explored the analysis and presentation through Natural Language Generation (NLG) of data from the Moodle Virtual Learning Environment[1], which students were using as part of a Distance Learning course.

Student involvement was incorporated at all stages of the project. The design, development, and testing phases of LARC were informed by formal student representation, motivated by a general concern for ethical practices in data collection. Students can experience learning analytics applied to them as individuals as "snooping" (Parr, 2014), and the LARC project aimed to avoid this by giving students a chance to interact with their data. The students taking part in the pilot project were studying either "Understanding Learning in the Online Environment" or "Digital Futures for Learning" and were asked to provide feedback about the LARC system. We intended that some of the generated texts would be controversial, and would provoke strong reactions from students, to cause them to consider aspects of data interpretation and ownership.

2 Related Work

Previous research has investigated the use of NLG techniques to generate reports from time-series data in a number of different domains (Sripada et al., 2003b). These include medical data summarization in the BabyTalk Project, providing decision support in a Neonatal Intensive Care Unit (Gatt et al., 2009; Hunter et al., 2011), and weather forecasts in SUMTIME-MOUSAM (Sripada et al., 2003a).

There are also a number of systems which have analysed data from Virtual Learning Environments, presenting it to the students themselves or to their institutions (Gašević et al., 2015), and global companies such as Civitas[2] and Knewton[3] offer large-scale data analytics solutions to educational institutions and publishers. However, there are only a few

[1] https://moodle.org

[2] https://www.civitaslearning.com/

[3] https://www.knewton.com/approach/platform

Proceedings of The 9th International Natural Language Generation conference, pages 207–211,
Edinburgh, UK, September 5-8 2016.

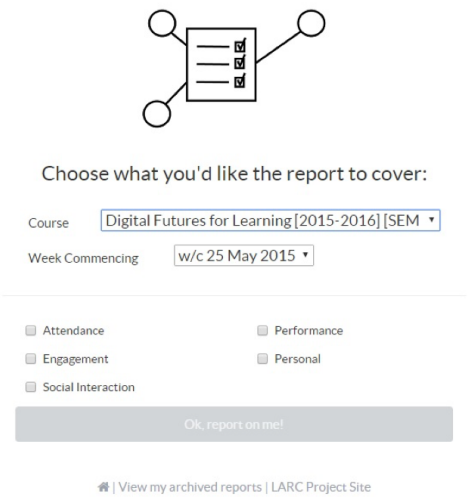

Figure 1: LARC user interface

systems which have made use of NLG in presenting their data.

The SkillSum project generated reports for adult students taking a basic skills test designed to check their basic numeracy and literary skills (Reiter et al., 2006; Williams and Reiter, 2008). It generated reports which used language tailored to the reading ability of the students, informing them whether or not their skill levels were suited to adult education courses in which they were interested. Our system deals with different sorts of data, and provides a more general report, rather than giving tailored advice on course choice.

The research of Gkatzia et al (2013) relates most closely to the LARC project - they generated reports based on time-series data from student lecture attendance and weekly questionnaires, and used reinforcement learning informed by the lecturers' method of providing feedback to choose the content to be contained in the report. They also compared students to their own past behaviours and to the student cohort. Our work differs from theirs in the nature of the data – the LARC data was all automatically gathered from the Moodle platform – and in the fact that we allow the students rather than the software to choose what should be presented. In addition, our pilot system was entirely rule-based. We also performed sentiment analysis on student forum posts, and to our knowledge are the first to have presented the results of this analysis using NLG.

> Your attendance has in general been excellent but this week you logged on less often than usual.
> You have mostly been very engaged with the course content and this week you seemed more interested in the topic than usual.
> You have usually been extremely social during the course but this week you interacted less with others than usual. Most of your forum posts were neutral in tone, some were positive, and none were negative.
> You are fairly concerned what others in the class think about you.
> You are in the middle third of students for social interaction and engagement, but the highest third of students for attendance.

Figure 2: LARC generated report with all 5 themes selected

3 User Interface

The LARC interface consists of a web form, accessible to students when they log in through a secure system. The interface is shown in Figure 1. The students used check boxes to select one or more of the five themes presented (described in Section 4), and the week for which they would like their report to be generated. They could generate a report as often as they wanted, and if they wished, they could at any time generate a report for a previous week. The data used to generate the LARC reports was automatically downloaded once a day from the Moodle server into an SQL database, which was then accessed and analysed by the NLG software in order to construct a report.

4 Report Themes

We chose five report themes, and values were set by the course lecturer in order to quantify the student's performance. For *attendance* (weekly login frequency), *engagement* (clicks on course pages) and *social* (accesses of the course discussion forum), (Table 1). For *personal*, we counted a student's posts to the course's "Introductions" forum and clicks by others on their profile. For *performance* we compared the student's performance to the average of the student cohort (Section 5.2). In addition, we generated a summary of sentiment analysis carried out on the student's posts to the discussion forum (Section 5.3).

5 Report Generation and Contents

The report consisted of a short paragraph on each of the selected themes, generated in real-time by a

Theme \ Rating	poor	adequate	good	excellent
attendance	<5	<10	<15	>=15
engagement	<10	<25	<100	>=100
social	<4	<8	<12	>=12

Table 1: Report Theme Values

Rating \ This Week vs Previous	**higher** (+)	**lower** (-)
poor (-)	contrast	similarity
adequate (-)	contrast	similarity
good (+)	similarity	contrast
excellent (+)	similarity	contrast

Table 2: Individual Comparisons

Java-based system with custom templates. Figure 2 shows a sample report.

5.1 Individual Comparisons

For the *attendance*, *engagement* and *social* themes, the student's scores for the week selected were compared to their average scores up to that point, and a sentence containing a comparison between the two was generated. The clause describing the student's general performance expresses the value judgments as described in Section 4, and in the generated texts the weekly score is compared to the previous average using Rhetorical Structure Theory similarity and contrast relations (Mann and Thompson, 1998) (Table 2). If the scores are identical, no comparison is included. A contrast relation is expressed by the conjunction "but" and a similarity relation by "and". For example, if the student's average attendance was 12 (good, +), and the current week 18 (higher, +), the generated sentence would be "Your attendance has in general been *good*, **and** this week you logged on *more* than usual" whereas if the average engagement was 120 (excellent +) and the current week 80 (lower, -) we would generate "You have mostly been *very engaged* with the course content **but** this week you seemed *less* interested in the topic than usual".

5.2 Cohort Comparisons

If a student selected the *performance* theme, we generated a sentence comparing their average performance to their course cohort. We included comparisons on *attendance*, *engagement* and *social* if any had been selected, or an average of all three if not.

We calculated the student's position within the cohort, and assigned them to the bottom, middle, or top third for each chosen theme. If more than one theme was chosen, we aggregated all of the matching positions, and combined the dissimilar ones with similarity or contrast relations, as in the following example "You are in the *highest* third of students for attendance **and** engagement, **but** the *lowest* third for social interaction."

5.3 Sentiment Analysis

To enable us to include a summary of the sentiments expressed in the students' forum posts, we experimented with two sentiment analysis packages. The first, part of the Stanford CoreNLP tools (Manning et al., 2014), comes with a model trained on movie review texts, which did not transfer well to our domain. Since we did not have any annotated data with which to train our own model, we used the rule-based Pattern system, (De Smedt and Daelemans, 2012) which generalized more successfully. We obtained sentiment subjectivity and polarity ratings for each blog post, which ranged between 1 and 0. For each post, we considered the sentiment to be neutral unless the subjectivity and polarity were both greater than .2. These levels were set after initial testing and are an aspect which we would hope to refine in future versions of the system (Section 7).

5.4 Quantifiers

There is a large body of research on the theory and use of quantifiers (Moxey and Sanford, 1986; Bos and Nissim, 2006; Lappin, 2000), and Varges and van Deemter (2005) give a theoretical handling of generation, but we are not aware of existing systems which actually generate quantifiers. We based our algorithm on recent research which investigated which quantifiers human subjects found acceptable when presented with an image of a bowl containing different numbers of blue and green candies (Yildirim et al., 2013; Yildirim et al., 2016). They found a high degree of individual variation and overlap but general consensus on some areas, having analysed human subjects' classification of "naturalness" for five quantifiers, and based on their results, we used the quantifiers shown in Table 3 to describe the results of our sentiment analysis (Section 5.3). For example, if a student made 20 blog posts, of which 13 (65%) were positive, 5 (25%) neutral and 2 (10%) negative the output would be "**Many** of your blog posts were positive, **some** were neutral and **few**

all	most	many	some	few	none
total	>60%	>40%	>20%	>1	0

Table 3: Generated Quantifiers

were negative".

5.5 Filler Sentences

If the student did not select all of the themes, filler sentences were added so that all reports would be of similar length. Each theme has a set of four sentences for each level of performance, so that if a single theme is selected, there will still be five sentences in the report. These sentences give general guidance, for example "attendance is key to achieving your aims on the course, and this an area you could improve upon" and "engaging with course content demonstrates your participation in the course, and you are showing yourself to be highly active".

6 Initial Student Reactions

Student feedback was given throughout the duration of the course, and as a result some changes were made before the trial ended, while others will be considered in future. This feedback is anecdotal and cannot be considered an evaluation, but we were informed by some comments and changed the structure of the output accordingly. Some feedback was positive, but we have concentrated here on comments which raised issues for us.

The initial version of the system presented only the average behaviour of the students over all of the preceding weeks, and therefore there was often no change in a student's report from one week to the next if their behaviour had remained consistent. We therefore introduced comparisons with the current week, to make it clear that the data was being analysed on a weekly basis.

Some students wanted to see the numbers underlying the generated sentences, so at the end of the pilot we introduced a data summary at the bottom of the report We intend to present this in a more user-friendly format, and integrate it with potential future graphical representations

One student commented that "As a student, I like friendly feedback" and wanted to see more "human language" for example encouraging comments such as "well done". Several students mentioned their worries about the ethics of learning analytics, with comments such as "We should adopt an ethical approach when extracting conclusions from analytical reports: they should be reviewed with caution" and one quotes EDUCAUSE[4] (a non-profit association whose stated mission is to "advance higher education through the use of information technology"), saying "Even then the best evaluative algorithms can result in misclassifications and misleading patterns, in part because such programs are based on inferences about what different sorts of data might mean relative to student success".

7 Future Work

As LARC was a pilot project, we did not have the time or resources for all of the development that we would have liked to carry out. We would like to add several functionalities to the system:

- Allow the students to choose from multiple report styles or personalities, which could be more encouraging, or more critical.
- Investigate more alternative sentiment analysis packages, and potentially allow the students to compare the outputs on their forum posts.
- Add graphical elements to the report. We would like to accompany the texts with visualizations such as circle graphs or heat maps in order to give a different view over the data.

We would also like to carry out formal user evaluations on various aspects of the generated texts:

- How the students rate the generated texts compared to a fixed baseline, and hand-written reports
- The use of the various quantifiers in describing the sentiment of forum posts
- The use of the contrast/similarity comparisons and the ordering of the various types of data within them

Finally and most importantly, we would like to continue our work to ensure that students understand how their data are used, and are happy with the resulting analyses.

[4] http://www.educause.edu

References

Johan Bos and Malvina Nissim. 2006. An Empirical Approach to the Interpretation of Superlatives. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 9–17, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *The Journal of Machine Learning Research*, 13(1):2063–2067.

Rebecca Ferguson. 2012. Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6):304.

Dragan Gašević, Shane Dawson, and George Siemens. 2015. Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1):64–71.

Albert Gatt, Francois Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur, and Somayajulu Sripada. 2009. From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. *Ai Communications*, 22(3):153–186.

Dimitra Gkatzia, Helen Hastie, Srinivasan Janarthanam, and Oliver Lemon. 2013. Generating student feedback from time-series data using Reinforcement Learning. *Proceedings of the 14th European Workshop on Natural Language Generation*.

James Hunter, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, Cindy Sykes, and Dave Westwater. 2011. Bt-nurse: computer generation of natural language shift summaries from complex heterogeneous medical data. *Journal of the American Medical Informatics Association*, 18(5):621–624.

Shalom Lappin. 2000. An intensional parametric semantics for vague quantifiers. *Linguistics and philosophy*, 23(6):599–620.

William Mann, C. and Sandra Thompson, A. 1998. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 3:243–281.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Linda M. Moxey and Anthony J. Sanford. 1986. Quantifiers and Focus. *Journal of Semantics*, 5(3):189–206, January.

Chris Parr. 2014. Lecturer calls for clarity in use of learning analytics. *Times Higher Education Supplement*, 6 Novebmer. https://www.timeshighereducation.com/news/lecturer-calls-for-clarity-in-use-of-learning-analytics/2016776.article.

Ehud Reiter, Sandra Williams, and Lesley Crichton. 2006. Generating feedback reports for adults taking basic skills tests. In *Applications and Innovations in Intelligent Systems XIII*. Springer.

Somayajulu Sripada, Ehud Reiter, and Ian Davy. 2003a. Sumtime-mousam: Configurable marine weather forecast generator. *Expert Update*, 6(3):4–10.

Somayajulu G. Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. 2003b. Generating English summaries of time series data using the Gricean maxims. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 187–196. ACM.

Sebastian Varges and Kees Van Deemter. 2005. Generating referring expressions containing quantifiers. *Proceedings of the 6th International Worskhop on Computational Semantics*.

Sandra Williams and Ehud Reiter. 2008. Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, 14(04).

Ilker Yildirim, Judith Degen, Michael K. Tanenhaus, and T. Florian Jaeger. 2013. Linguistic variability and adaptation in quantifier meanings. In *Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society*, pages 3835–3840.

Ilker Yildirim, Judith Degen, Michael K. Tanenhaus, and T. Florian Jaeger. 2016. Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, 87:128–143.