Institution: University of Aberdeen

Unit of Assessment: 11 (Computer Science and Informatics)

Title of case study: Data2Text

1. Summary of the impact

Data-to-text utilises Natural Language Generation (NLG) technology that allows computer systems to generate narrative summaries of complex data sets. These can be used by experts, professional and managers to better, and quickly, understand the information contained within large and complex data sets. The technology has been developed since 2000 by Prof Reiter and Dr Sripada at the University of Aberdeen, supported by several EPSRC grants. The Impact from the research has two dimensions.

As *economic* impact, a spinout company, Data2Text (www.data2text.com), was created in late 2009 to commercialise the research. As of May 2013, Data2Text had 14 employees. Much of Data2Text's work is collaborative with another UK company, Arria NLG (www.arria.com), which as of May 2013 had about 25 employees, most of whom were involved in collaborative projects with Data2Text.

As impact on *practitioners and professional services*, case studies have been developed in the oil & gas sector, in weather forecasting, and in healthcare, where NLG provides tools to rapidly develop narrative reports to facilitate planning and decision making, introducing benefits in terms of improved access to information and resultant cost and/or time savings. In addition the research led to the creation of *simplenlg* (http://simplenlg.googlecode.com/), an open-source software package which performs some basic natural language generation tasks. The *simplenlg* package is used by several companies, including Agfa, Nuance and Siemens as well as Data2Text and Arria NLG.

2. Underpinning research

Data-to-text technology was developed by Professor Reiter, Dr Sripada (originally Research Fellow, now Senior Lecturer), and Professor Jim Hunter, at the University of Aberdeen from 2000. The work arose from Reiter's interest in "language and the world", he wanted to explore how real-world information is mapped onto language, and focused on situations where human authors wrote English-language narrative summaries of complex numeric data sets. Although the initial motivation was basic research in computer and cognitive science, it became apparent that there was considerable commercial interest in software which could automate the task of writing such summaries.

The research began with the EPSRC project *SumTime: Generating Summaries of Time-Series Data* (2000-2003) [GR/M76881/01]. Reiter and Hunter were investigators, with Sripada employed as a research fellow. The primary focus of *SumTime* was to generate automated narrative weather forecasts from numerical weather prediction data, although the group also worked on summarising data from gas turbines, and from electronic patient records in a hospital. One striking scientific finding were the large differences in words and language used by human weather forecasters. Another major outcome was that an evaluation of the automated weather forecasts generator (which was operationally deployed at the Aberdeen office of Weathernews) showed that many forecast users preferred computer-generated texts over human-written forecasts, in part because of greater consistency in their language and word use. *SumTime* was followed by an EPSRC CASE studentship (Ross Turner, supervised by Sripada, by then a lecturer at Aberdeen, and Reiter), which developed *Roadsafe*, a more sophisticated system which generated weather forecasts over a spatial region (instead of just at one point); this was in conjunction with another company, Aerospace and Marine International. This project developed major advances in theory and algorithms for generating spatial reference. The *SumTime* and *Roadsafe* work formed the basis of commercial weather-forecasting work done by Data2Text, which now employs Dr Turner.

After spending time on an ESRC-funded Paccit-Link project (*Automatic Generation of Personalised Basic Skills Summary Reports*) [L328253023], which summarised education assessment data for adults with poor basic skills, the group then embarked on the *Babytalk* project (2006-2012), supported by three EPSRC grants [EP/D049520/1,EP/D05057X/1,EP/H042938/1] and two EPSRC DTA studentships; this was a collaboration with NHS Lothian. Reiter, Hunter, and Sripada were investigators on the main grants. *Babytalk*'s goal was to develop software which could automatically generate summaries of clinical data about premature infants in neonatal intensive care for use by doctors, nurses and parents. *Babytalk* was much more challenging than *SumTime*, because the data was more complex (it included event data such as medical interventions, as well as time-series data, and also required extracting information from free text). The summaries also had to be generated, using the same underlying architecture, for three very different audiences. Probably the most important scientific finding of *Babytalk* was that it proved the task could be done: it was possible to build, and deploy in a hospital environment, a data-to-text system which could generate useful summaries of a complex heterogeneous data set for diverse audiences. Another result of this work was the development of an architecture and software framework that could be used to generate summaries of heterogeneous data sets across a range of potential application areas. Thus *Babytalk* provided a basis for Data2Text's work in the oil & gas industry.

Babytalk also allowed the group to refine an open source software product *simplenlg* into a commercially valuable form. Earlier versions of simplenlg had been used in Aberdeen since the late 1990s, but during Babytalk simplenlg was much improved (in robustness and documentation as well as functionality), and then released on an open-source basis for free use for both academic research and commercial applications.

The group also did some work on applying data-to-text technology to assist people with disabilities; this work was supported by 2 EPSRC grants [EP/F066880/1,EP/H022376/1]. The *How Was School Today* project, which helped non-speaking children write stories about their school day, was the subject of an EPSRC Impact study and also mentioned in several EPSRC reports.

3. References to the research

[R1] E Reiter and S Sripada (2002). Human Variation and Lexical Choice. *Computational Linguistics* **28**:545-553. http://dx.doi.org/10.1162/089120102762671981

[R2] E Reiter, S Sripada, J Hunter, J Yu, and I Davy (2005). Choosing Words in Computer-Generated Weather Forecasts. *Artificial Intelligence* **167**:137-169. http://dx.doi.org/10.1016/j.artint.2005.06.006

** These papers describe key findings from the SumTime project. The first focuses specifically on differences in word usage between different authors, looking at several domains. The second focuses on the weather forecast generation domain, including a more detailed analysis of differences in word and language usage between different weather forecasters, and an evaluation which showed that forecast users in some cases preferred SumTime texts to texts written by human forecasters. *Papers [R1] and [R2] above best indicate the quality of underpinning research.*

[R3] A Law, Y Freer, J Hunter, R Logie, N McIntosh, J Quinn (2005). A Comparison of Graphical and Textual Presentations of Time Series Data to Support Medical Decision Making in the Neonatal Intensive Care Unit. *Journal of Clinical Monitoring and Computing* **19**:183-194. http://dx.doi.org/10.1007/s10877-005-0879-3

** This paper, which laid the foundation for much of the Babytalk work, shows that doctors and nurses make better decisions from textual summaries of clinical data than from visualisations, at least in some contexts.

[R4] F Portet, E Reiter, A Gatt, J Hunter, S Sripada, Y Freer, C Sykes (2009). Automatic Generation of Textual Summaries from Neonatal Intensive Care Data. *Artificial Intelligence* **173**:789-816. [Sripada1 in the REF2 for this unit.]

[R5] J Hunter, Y Freer, A Gatt, E Reiter, S Sripada, C Sykes (2012). Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial Intelligence in Medicine* **56**:157–172. [Reiter2 in the REF2 for this unit.]

** These papers describe key findings of the Babytalk doctor and nurse systems (the parent system is still being evaluated), including architecture and on-ward evaluation of the nurse system. *Paper [R4] above best indicates quality of underpinning research.*

[R6] R Black, A Waller, R Turner, E Reiter (2012). Supporting Personal Narrative for Children with Complex Communication Needs. *ACM Transactions on Computer-Human Interaction* **19(2)**, Article 15. [Reiter4 in the REF2 for this unit.]

** This paper describes key findings of the *How Was School Today?* project.

4. Details of the impact

Following interest in the results of the research, in late 2009 the decision was taken to create a spinout company, Data2Text, to commercialise the underpinning research. Data2Text essentially develops bespoke data-to-text software applications for large organisations, based on a generic data-to-text software library. The technology and expertise developed in *SumTime* and *BabyTalk* are very much at the core of Data2Text's activities. In addition to its commercial contracts (which cannot be described in detail here because of commercial confidentiality [S2,S3]), Data2Text was awarded a Smart Award from Scottish Enterprise. The company currently employs 14 staff and has a turnover of approximately £1M/year. In 2012 Data2Text formed a partnership with Arria NLG, who acquired a minority shareholding in Data2Text [S1]. In October 2013 Arria NLG acquired Data2Text, and in November 2013 an application was made to the Alternative Investment Market (AIM) through the London Stock Exchange for an Initial Public Offering for shares in Arria NLG Plc. The expected size of this offer was £6.1M [S7], with a likely valuation of £102M. Admission to AIM is expected in early December.

Beyond the economic impact of the spin out company creation, NLG technology is also having an impact both economically and on practitioners and professional services through commercial partnerships between Data2Text/Arria NLG and other organisations. The Arria NLG website refers to a number of case studies [S6], and refers strongly to the impact derived from the original research at the University of Aberdeen. Also see the description of the November 2013 Initial Public Offering on the London Stock Exchange [S7], which states that "[t]he Group's core product is known as the Arria NLG Engine, which originates from research at the University of Aberdeen". NLG products are currently developed through Data2Text and Arria NLG in two key application areas: oil & gas and weather forecasting. The companies are also exploring applications in financial services and healthcare.

In oil & gas, NLG applications have been used to monitor alerts produced by rotating equipment on oil platforms. Operating continuously through a 365 day cycle, breakdown of equipment results in lost production time for oil & gas operators. A Data2text/Arria NLG system is being used by a multinational oil company to automatically produce situation analyses, based on data streams from turbines, compressors, pumps, generators and engines, when a surveillance alert is triggered [S2]. Manually writing such analyses can take an experienced engineer several hours; the NLG software does it in minutes. As mentioned above, this system is inspired by the *Babytalk* research project; essentially *Babytalk*-inspired ideas are used to monitor equipment on oil platforms instead of premature infants. This has current and potential economic impacts on the multinational oil company partner, Shell. Shell has stated "by adding the Arria NLG Engine to traditional surveillance technologies to monitor their global rotating equipment assets, a one percentage point uptick in production uptime could be achieved". Further, "if the Arria NLG Engine can be added to all of their global production assets […this…] could equate to billions of dollars of increased production per year" [S2,S6]. The description of Arria NLG in the IPO made in November 2013 states that the software "is being used to analyse the performance data of large scale industrial machinery located on […] Oil & Gas platforms in the Gulf of Mexico, producing real-time written reports for engineers at the […] centre for surveillance of offshore operations."

Application of NLG technology in weather forecasting began as part of the research programme in 2000-9 with the Aberdeen offices of Weathernews and Aerospace and Marine International. Since 2009, this has extended to collaboration with a leading weather service, the Met Office. To date the NLG technology has been used to generate site specific, on-demand, detailed weather forecasts. The technology is capable of preparing detailed 3 day weather forecasts for 5,000 different locations in less than one minute, the equivalent for a human forecaster would require one and a half months. "Right now, our NLG software tackles tasks that would be impossible for forecasters to complete manually. It would take a forecaster 1.5 months to create the equivalent of our system's one-minute output" (detailed, 3-day weather forecasts for 5,000 different locations) [S3,S6]. In this way, the Met Office can offer additional services to a wide range of customers.

Researchers at the University of Aberdeen also created and released the open source data-to-text resources *simplenlg*. *simplenlg* is a Java software library (currently in version 4.4) for doing some

Natural Language Generation processing (surface realisation and a small amount of microplanning) [S8]. It was initially released purely for research use, but has been updated with significant enhancements and documentation since 2010 (from version 4.0) to be also available for commercial use on an open-source basis (simplenlg.googlecode.com). Since the release of version 4.2 (April 2011), almost 2,000 copies of *simplenlg* have been downloaded, and an indicator of the increasing size of the user community is the increase in download statistics through these versions: version 4.2, 228; version 4.3, 733; and version 4.4, 1032. The *simplenlg* library is currently being used by a number of commercial companies (as well as academic groups) around the world, especially in healthcare-related applications. A good example of an Open Source App that uses *simplenlg* is the Augmentative and Alternative Communication (AAC) Speech Communicator, “an Android application for people with speech disabilities that forms sentences from a list of pictograms” [S9]. According to Google Play this App has been installed over 10,000 times, and has an aggregate rating of 4/5 (25 reviews). Good examples of how *simplenlg* has been integrated into products include Agfa, where it is used in a cardiology clinical reporting application [S5], and Siemens, where it is used it in healthcare software sold to US hospitals [S4]. It is also being used commercially for assistive technology; for example Technabling (another Aberdeen Computing Science spinout) use it in their portable sign language translator [S10], which is to be launched as a product in Autumn 2013.

5. Sources to corroborate the impact

[S1] President, Arria NLG – will corroborate the partnership between Data2Text and Arria NLG, the integration of underpinning research in the Arria NLG engine, and economic impact (including staff, turnover and acquisition).

[S2] Senior Surveillance Engineering Specialist, Shell – will corroborate the economic impact and impact on practice due to the use of Data2Text technologies.

[S3] Head of Customer Applications, Met Office – will corroborate the economic impact and impact on practice due to the use of Data2Text technologies.

[S4] Software Developer, Siemens Medical Solutions – will corroborate the impact on practice due to the use of Data2Text technologies.

[S5] Software Engineer, Agfa Healthcare – will corroborate the impact on practice due to the use of Data2Text technologies.

[S6] Arria NLG case studies: https://www.arria.com/case-studies-A230.php – will corroborate breadth of sectors in which underpinning research is applied through the Arria NLG engine.

[S7] Initial Public Offering of Arria NLG: http://www.londonstockexchange.com/exchange/prices-and-markets/stocks/new-and-recent-issues/new-recent-issue-details.html?issueId=8923 - corroborates the fact that the impact originates from research at the University of Aberdeen, the acquisition of Data2Text by Arria NLG and the IPO being issued, the deployment of the software on Oil & Gas platforms in the Gulf of Mexico, and the size of the economic impact made through the commercialisation of this research.

[S8] Simplenlg website: http://simplenlg.googlecode.com/ (includes link to simplenlg discussion group: http://groups.google.com/group/simplenlg) – will corroborate size of user community and engagement of that community in using simplenlg in practice.

[S9] The AAC Speech Communicator (Open Source App that uses *simplenlg*): http://aacspeech.org/ - will corroborate the use of simplenlg in this App.

[S10] The portable sign language translator: http://www.pslt.org/ - will corroborate a commercial use of simplenlg.