The Importance of Narrative and Other Lessons from an Evaluation of an NLG System that Summarises Clinical Data

Ehud Reiter, Albert Gatt, François Portet
Dept of Computing Science
University of Aberdeen, UK
{e.reiter,a.gatt,fportet}@abdn.ac.uk

Marian van der Meulen*
Dept of Psychology
University of Edinburgh, UK
m.a.van-der-meulen@sms.ed.ac.uk

Abstract

The BABYTALK BT-45 system generates textual summaries of clinical data about babies in a neonatal intensive care unit. A recent task-based evaluation of the system suggested that these summaries are useful, but not as effective as they could be. In this paper we present a qualitative analysis of problems that the evaluation highlighted in BT-45 texts. Many of these problems are due to the fact that BT-45 does not generate good narrative texts; this is a topic which has not previously received much attention from the NLG research community, but seems to be quite important for creating good data-to-text systems.

1 Introduction

Data-to-text NLG systems produce textual output based on the analysis and interpretation of non-linguistic data (Reiter, 2007). Systems which produce short summaries of small amounts of data, such as weather-forecast generators (Reiter et al., 2005), have been one of the most successful applications of NLG, and there is growing interest in creating systems which produce longer summaries of larger data sets.

We have recently carried out an evaluation of one such system, BT-45 (Portet et al., 2007), which generates multi-paragraph summaries of clinical data from a Neonatal Intensive Care Unit (NICU). The summaries cover a period of roughly 45 minutes, and describe both sensor data (heart rate, blood oxygen saturation, etc, sampled at 1 sec intervals) as well as discrete events such as drug administration; they are intended to help medical staff make treatment decisions. This evaluation showed that from a decision-support perspective, the BT-45 texts were as effective as visualisations of the data, but less effective than human-written textual summaries.

In addition to quantitative performance data, which is presented elsewhere (van der Meulen et al., submitted), the evaluation also gave us valuable clues about what aspects of data-to-text technology need to be improved in order to make texts generated by such systems more effective as decision support aids; this is the subject of this paper. Somewhat to our surprise, many of the problems identified in the evaluation relate to the fact that BT-45 could not produce a good narrative describing the data. Generation of non-fictional narratives is not something which has been the focus of much NLG research in the past, but our results suggest it is important, at least in the context of producing texts which are effective decision-support aids.

1.1 Background: Data-to-Text

Data-to-text systems are motivated by the belief that (brief) linguistic summaries of datasets may in some cases be more effective than more traditional presentations of numeric data, such as tables, statistical analyses, and graphical visualisations (even simple visual/graphical displays require relatively complex cognitive processing (Carpenter and Shah, 1998)). Also linguistic summaries can be delivered in some contexts where visualisations are not possible, such as text messages on a mobile phone, or when the user is visually impaired (Ferres et al., 2006). In the NICU domain, Law et al. (2005) conducted an experiment which showed that medical professionals were more likely to make the correct treatment deci-

*Now at the Department of Clinical Neurosciences, University Hospital, Geneva, Switzerland

sion when shown a human-written textual summary of the data than when they were shown a graphical visualisation of the data.

A number of data-to-text systems have been developed and indeed fielded, especially in the domain of weather forecasts (Goldberg et al., 1994; Reiter et al., 2005). Most of these systems have generated short (paragraph-length or smaller) summaries of relatively small data sets (less than 1KB). Some research has been on systems that summarise larger data sets (Yu et al., 2007; Turner et al., 2008), but these systems have also generated paragraph-length summaries; we are not aware of any previous research on generating multi-paragraph summaries in a data-to-text system.

Data-to-texts systems have been evaluated in a number of ways, including human ratings (the most common technique) (Reiter et al., 2005), BLEU-like scores against human texts (Belz and Reiter, 2006), post-edit analyses (Sripada et al., 2005), and persuasive effectiveness (Carenini and Moore, 2006). However, again to the best of our knowledge no previous data-to-text system has been evaluated by asking users to make decisions based on the generated texts, and measuring the quality of these decisions.

2 BabyTalk and BT-45

Law et al. (2005) showed that human-written textual summaries were effective decision-support aids in NICU, but of course it is not practical to expect medical professionals to routinely write such summaries, especially considering that the summaries used by Law et al. in some cases took several hours to write. The goal of the BABYTALK research project is to use NLG and data-to-text technology to automatically generate textual summaries of NICU data, for a variety of audiences and purposes. The first system developed in BABYTALK, and the subject of this paper, is BT-45 (Portet et al., 2007), which generates summaries of 30-60 minute chunks of clinical data, for the purpose of helping nurses and doctors make appropriate treatment decisions.

An example of BABYTALK input data is shown in Figures 1 (sensor data) and 2 (selected event data). Figure 3 shows the human-written corpus text for this scenario, and Figure 4 shows the BT-45 text generated for this scenario. Note that for the purposes of

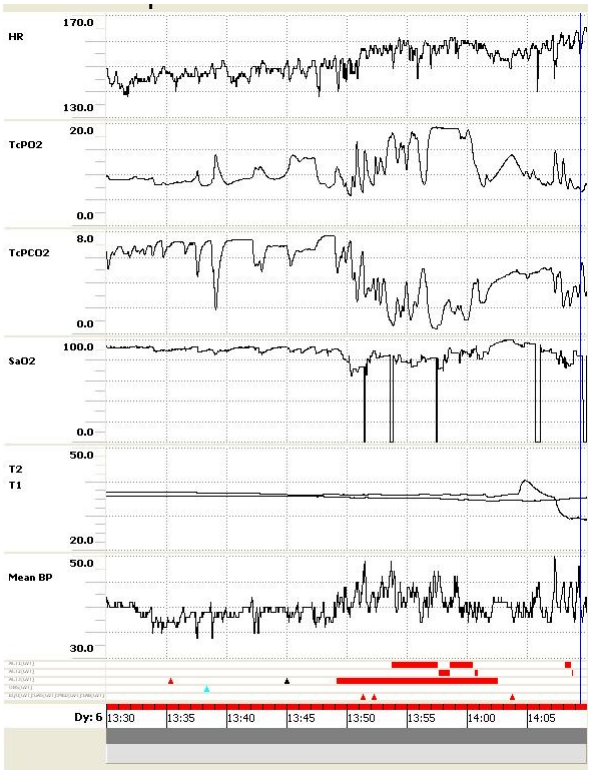

Figure 1: Example Babytalk Input Data: Sensors
HR = Heart Rate; TcPO2 = blood O2 level; TCPCO2 = blood CO2 level; SaO2 = oxygen saturation; T1 = chest temperature; T2 = toe temperature; mean BP = blood pressure. The bars and triangles at the bottom show the time of discrete events (Figure 2).

event	*time*
Blood transfusion	13.35
Intermittent crying	13.38
FiO2 (oxygen level) changed to 50%	13.51
Incubator temperature changed to 36.4	13.52
Attempt to insert line	13.53
Line removed	13.57
Attempt to insert line	13.58
Line removed	14.00
FiO2 (oxygen level) changed to 45%	14.03
Attempt to insert line	14.08
Line removed	14.09

Figure 2: Example Babytalk Input Data: Selected Discrete Events

You saw the baby between 13:30 and 14:10.

Over the course of the monitoring period the HR increases steadily from 140 to 165; the BP varies between 41 and 49.

At the start of the period T1 is 37 and T2 is 35.8C.

During the first 15 minutes the pO2 is 7.8-9.2 and the pCO2 is 5.9-7.3.

At 13:35 a blood transfusion is commenced.

At 13:38 the baby is crying and there are a few upward spikes in the pO2 trace corresponding to downward spikes in the pCO2 trace. At 13.45 the humidity on the incubator walls is wiped away and T1 and T2 fall to 36.3 and 35.4 respectively

At 13:50 the baby is examined. There is a desaturation to 72% and the FiO2 is changed to 50%. Between now and 14.10 there are several attempts to site a peripheral cannula. The pO2 and pCO2 both have spikes in the traces, pO2 reaching 19.2 and pCO2 reaching 0.4. There are several episodes of artefact in the oxygen saturation trace.

T1 and T2 fall to 36.2 and 35.7 and the oxygen saturation falls to 65%. The FiO2 is adjusted to 50%. Also at this time the incubator temperature is adjusted to 36.4C.

At 14:03 with the pO2 at 13.9 and oxygen saturation at 100%, the FiO2 is reduced to 45

At 14:04 T1 rises sharply to 40, then drops fairly steeply to 28.5C. Between 14.06 and 14.10 there are several spikes in the pO2 and pCO2 traces but by 14.10 the pO2 is 8, pCO2 is 3.7, the oxygen saturation is 71%, the HR is 163, the BP 39, T1 29C and T2 35.4C.

Figure 3: Corpus Text for Fig 1, 2 data

You saw the baby between 13:30 and 14:10. Heart Rate (HR) = 149. Core Temperature (T1) = 37.0. Peripheral Temperature (T2) = 35.8. Transcutaneous Oxygen (TcPO2) = 9.5. Transcutaneous CO2 (TcPCO2) = 6.7. Mean Blood Pressure (mean BP) = 41. Oxygen Saturation (SaO2) = 94.

Over the next 39 minutes SaO2 decreased to 81, T2 decreased to 34.5, HR stayed at around 151, T1 decreased to 36.2 and mean BP stayed at around 40.

A blood transfusion was given to the baby at 13:35.

At 13:50 there was a desaturation down to 65. As a result, Fraction of Inspired Oxygen (FIO2) was set to 50

There were 3 failed attempts to insert a peripheral venous line at 13:53. TcPO2 suddenly decreased to 8.1. SaO2 suddenly increased to 92. TcPO2 suddenly decreased to 9.3. There was a spike in TcPO2 up to 14.8. There had been another spike in T1 up to 40.5. FIO2 had been lowered to 45%. Previously the baby had been examined.

Figure 4: BT-45 Text for Fig 1, 2 data

this paper, we have deliberately selected a relatively poor quality BT-45 output text.

The processing performed by BT-45 is described by Portet et al. (2007). Very briefly, BT-45 generates texts in four stages:

- *signal analysis*: for example, detecting spikes in the sensor channels; this is done using adaptations of standard pattern detection and recognition algorithms.
- *data abstraction*: for example, identifying three line-insertion-attempt and line-removal events within a short span of time and grouping these into the higher-level concept LINE-INSERTION-PROCESS-FAILED (BT-45 includes a domain ontology of such concepts); this is done using knowledge-based techniques.
- *document planning*: for example, deciding not to mention most of the spikes in O2 and CO2; this is primarily done in a bottom-up fashion, using information (computed by the data abstraction module) on the medical importance of events, and also on causal and other relationships between events.
- *microplanning and realisation:* producing the actual text shown in Figure 4; this is mostly done using relatively standard NLG techniques, although we have developed new techniques for communicating temporal information and relationships.

3 Evaluation of BT-45

BT-45 was evaluated by asking medical staff to make decisions about what actions they should take with regard to a baby, after viewing either a BT-45 text, a human-written textual summary, or a visualisation of the baby's data; this was similar in general terms to the experiment described by Law et al. (2005). van der Meulen et al. (submitted) gives de-

tails about the evaluation design and quantitative results of the evaluation; in this paper we just briefly summarise these aspects of the evaluation

Material: Our medical collaborators selected 24 scenarios (data sets), and defined 18 types of actions. For each of the data sets, they specified which of the 18 actions were appropriate, inappropriate, or neutral (neither appropriate nor inappropriate); one appropriate action was identified as the main target action. For the data set shown in Figures 1 and 2, for example:

- *Main target action*: Adjust monitoring equipment
- *Other appropriate actions*: calm/comfort baby, manage temperature, analyse blood sample
- *Neutral actions*: adjust CPAP (ventilation) settings, baby care (e.g., nappy change)
- *Inappropriate actions*: all other actions (e.g. blood transfusion, order X-Ray) (12 in all)

For each scenario, we created three presentations: a visualisation (similar to Figure 1), a human-written text summary written by our collaborators, and the summary produced by BT-45. Our collaborators were asked not to include any explicit medical interpretation of the data in their human-written summaries. For each scenario, our collaborators also prepared a text which gave background information about the baby.

When developing BT-45, we had access to the data collection that the scenario data sets were taken from (which includes several months of data), but did not know ahead of time which specific scenarios would be used in the experiment.

Subjects: 35 medical professionals, including junior nurses, senior nurses, junior doctors, and senior doctors.

Procedure: Each subject was shown 8 scenarios in each condition (visualisation, human text, BT-45 text) in a Latin Square design; all subjects were also shown the background texts. Subjects were asked to specify which actions should be taken for this baby, selecting actions from a fixed set of check-boxes; they were given three minutes to make this decision. Subjects were not explicitly asked for free-text comments, but any comments spontaneously made by subjects were recorded. Subject responses were scored by computing the percentage of appropriate actions they selected, and subtracting from this the percentage of inappropriate actions.

Results: The highest score was achieved by the human texts (mean score of 0.39); there was no significant difference between the BT-45 texts (mean score of 0.34) and the visualisations (mean score of 0.33). The largest differences occurred in the junior nurses group. van der Meulen et al. (submitted) present a detailed statistical analysis of the results.

Discussion: This shows that BT-45 texts were as effective as visualisation, but less effective than the human texts. This suggests that data-to-text technology as it stands could be useful as a supplement to visualisations (since some individuals do better with texts and some with graphics; also some data sets are visualised effectively and some are not), and in contexts where visualisation is not possible. But it also suggests that if we can improve the technology so that computer-generated texts are as effective as human texts, we should have a very effective decision-support technology.

4 Quantitative Comparison of BT-45 and Corpus Texts

In addition to the task-based evaluation described above, we also quantitatively compared the BT-45 and human texts, and qualitatively analysed problems in the BT-45 texts. Quantitative comparison was done by annotating the BT-45 and human texts to identify which events they mentioned. For each scenario, we computed the MASI coefficient (Passonneau, 2006) between the set of events mentioned in the BT-45 and human texts. The average MASI score was 0.21 ($SD = 0.13$), which is low; this suggests that BT-45 and the human writers choose different content. We also checked whether similar human and BT-45 texts (as judged by MASI score) obtained similar evaluation scores; in fact there was no significant correlation between MASI similarity of human and BT-45 texts and the difference between their evaluation scores.

We performed a second analysis based on comparing the structure (e.g., number and size of paragraphs) of the BT-45 and human texts, using a tree-edit-distance metric to compare text structures.

Again this showed that there were large differences between the structure of the BT-45 and human texts, and that these differences did correlate with differences in evaluation scores.

In other words, simple metrics of content and structural differences do not seem to be good predictors of text effectiveness; this is perhaps not surprising given the complexity of the texts and the information they are communicating.

5 Qualitative Analysis of Problems in BT-45 texts

The final step in our evaluation was to qualitatively analyse the BT-45 texts and the results of the task-based evaluation, in order to highlight problems in the BT-45 texts. Of course we were aware of numerous ways in which the software could be improved, but the evaluation gave us information about which of these mattered most in terms of overall effectiveness. We report this analysis below, including issues identified from subjects' comments, issues identified from scenarios where BT-45 texts did poorly, and problems identified via manual inspection of the texts. We do not distinguish between 'linguistic' and 'reasoning' problems, in part because it is usually difficult (and indeed somewhat artificial) to separate these aspects of BT-45.

5.1 Problems Identified by Subjects

Subjects made a number of comments during the experiment. Two aspects of BT-45 were repeatedly criticised in these comments.

5.1.1 Layout and bullet lists

Subjects wanted better layout and formatting, in the human texts as well as the BT-45 texts (BT-45 texts do not currently include any visual formatting). In particular, they wanted bullet lists to be used, especially for lab results. Such issues have been extensively discussed by other researchers (e.g., (Power et al., 2003)), we will not further discuss them here.

5.1.2 Continuity

BT-45 sometimes described changes in signals (or other events) which didn't make sense because they omitted intermediate events. For example, consider the last paragraph in the BT-45 text shown in Figure 4 (with italics added):

> There were 3 failed attempts to insert a peripheral venous line at 13:53. *TcPO2 suddenly decreased to 8.1*. SaO2 suddenly increased to 92. *TcPO2 suddenly decreased to 9.3*. There was a spike in TcPO2 up to 14.8. There had been another spike in T1 up to 40.5. FIO2 had been lowered to 45%. Previously the baby had been examined.

Subjects complained that it made no sense for TcPO2 to decrease to 9.3 when the last value mentioned for this parameter was 8.1

In this case (and in many others like it), BT-45 had identified the decrease events as being medically important, but had not assigned as much importance, and hence not mentioned, the increase event (TcPO2 went up to 19) between these decrease events. This is partially because BT-45 believed that a TcPO2 of 19 is a sensor artefact (not a real reading of blood oxygen), since 19kPa is a very high value for this channel. In fact this is a correct inference on BT-45's part, but the text is still confusing for readers.

We call this problem *continuity*, making an analogy to the problems that film-makers have in ensuring that scenes in a film (which maybe shot in very different times and locations) fit together in the eyes of the viewer. It is interesting to note that some of the human texts also seem to have continuity problems (for example, the text in Figure 3 says T2 falls to 35.4, and then says T2 falls to 35.7), but none of the subjects complained about continuity problems in the human texts. So some kinds of continuity violations seem more problematical to readers than others. Perhaps this depends on the proximity of the events both in the document structure and in time; we hope to empirically explore this hypothesis.

Continuity is just one aspect of the broader problem of deciding which events need to be explicitly mentioned in the text, and which can be omitted. Making such decisions is perhaps one of the hardest aspects of data-to-text.

5.2 Scenarios Where BT-45 did Badly

When analysing the results of the experiment, we noticed that BT-45 texts did as well as the human texts for scenarios based on five of the eight target actions; however they did significantly worse than

main target action	*human*	*BT-45*	*diff*
Adjust CPAP	0.37	0.37	0
Adjust monitoring equip	0.59	0.22	0.37
Adjust ventilation	0.22	0.23	-0.01
Extubate	0.14	0.12	0.02
Manage temperature	0.55	0.33	0.22
No action	0.61	0.43	0.18
Suction	0.34	0.42	-0.08
Support blood pressure	0.45	0.55	-0.10

Table 1: Average evaluation score by main target action

the human texts on the scenarios based on the other three actions (Adjust Monitoring Equipment, Manage Temperature, and No Action). Details are shown in Table 1; an ANOVA confirms that there is a significant effect of main target action on scores ($p < .001$). We have identified a number of reasons why we believe this is the case, which we discuss below.

5.2.1 Too much focus on medically important events

Content-selection in BT-45 is largely driven by rules that assess the medical importance of events and patterns. In particular, BT-45 tends to give low importance to events which it believes are due to sensor artefacts. While this strategy makes sense in many cases, it leads to poor performance in scenarios where the target action is Adjust Monitoring Equipment, when sensor problems need to be pointed out to the reader.

This can be seen in the example scenario used in this paper. The TcPO2 and TcPCO2 traces shown in Figure 1 are full of sensor artefacts (such as the implausibly high values of TcPO2 mentioned above). The human text shown in Figure 3 explicitly mentions these, for example (italics added)

> At 13:50 the baby is examined. There is a desaturation to 72% and the FiO2 is changed to 50%. Between now and 14.10 there are several attempts to site a peripheral cannula. *The pO2 and pCO2 both have spikes in the traces, pO2 reaching 19.2 and pCO2 reaching 0.4. There are several episodes of artefact in the oxygen saturation trace*.

The BT-45 text shown in Figure 4, in contrast, only mentions one spike in TcPO2, and does not mention any artefacts.

This is a difficult problem to solve, because in a context where medical intervention was needed, BT-45 would be correct to ignore the sensor problems. One solution would be for BT-45 to perform a top-level diagnosis itself, and adjust its texts based on whether it believed staff should focus on medical intervention or adjusting sensors. Whether this is desirable or even feasible is unclear; it relates to the more general issue of how a data-summarisation system such as BT-45 should be integrated with the kind of diagnosis systems developed by the AI/Medicine community.

5.2.2 Poor description of related channels

BT-45 essentially describes each channel independently. For temperature, however, it is often better to describe the two temperature channels together and even contrast them, which is what the human texts do; this contributes to BT-45's poor performance in Manage Temperature scenarios.

For example, in one of the Manage Temperature scenarios, the BT-45 text says

> Core Temperature (T1) = 36.4. Peripheral Temperature (T2) = 34.0. ...
> *(new paragraph)* Over the next 44 minutes T2 decreased to 33.4.

The human text says

> He is warm centrally but T2 drifts down over the 45 minutes from 34 to 33.3C.

The information content of the two texts is quite similar, but the human text describes temperature in an integrated fashion. Similar problems occur in other scenarios. In fact, over the 24 scenarios as a whole, the human texts include only three paragraphs which mention just one of the temperatures (T1 or T2, but not both), while the BT-45 texts include 18 such paragraphs.

BT-45's document planner is mostly driven by medical importance and causal relationships; although it does try to group together information about related channels, this is done as a secondary optimisation, not as a primary organising principle. The human texts place a much higher priority on

grouping 'physiological systems' (to use NICU terminology) of related channels and events together, including the respiratory and cardiac systems as well as the temperature system. We suspect that BT-45 should place more emphasis on systems in its document planning.

5.2.3 Poor long-term overview

BT-45 does not do a good job of summarising a channel's behaviour over the entire scenario. This isn't a problem in eventful scenarios, where the key is to describe the events; but it does reduce the effectiveness of texts in uneventful scenarios where the main target action is No Action (i.e., do nothing).

This problem can be seen in the text extracts shown in the previous section. Even at the level of individual channels, *He is warm centrally* is a better overview than *Core Temperature (T1) = 36.4*; and *T2 drifts down over the 45 minutes from 34 to 33.3C* is better than *Peripheral Temperature (T2) = 34.0. ... Over the next 44 minutes T2 decreased to 33.4.*

At a signal analysis level, BT-45 also does not do a good job of detecting patterns (such as spikes) with a duration of minutes instead of seconds. This contributes to the system's poor performance in Manage Temperature scenarios, because temperature changes relatively slowly.

We believe these problems can be solved, by putting more emphasis on analysis and reporting of long time-scale events in the BT-45 modules.

5.3 Other Problems

We manually examined the texts, looking for cases where the BT-45 texts did not seem clear. This highlighted a number of additional issues.

5.3.1 Describing events at different temporal time-scales

BT-45 does not always do a good job of correctly identifying long-term trends in a context where there are also short-term patterns such as spikes. In fact accurately detecting simultaneous events at different time-scales is one of the major signal analysis challenges in BT-45. There are linguistic issues as well as signal analysis ones; for example, should long-duration and short-duration events be described in separate paragraphs?

5.3.2 Poor communication of time

BT-45 texts often did not communicate time well. This is for a number of reasons, of which the most fundamental is problems describing the time of long-duration events. For instance, in our example scenario, the sequence of insert/remove line events in Figure 2 is analysed by the data-abstraction module as the abstract event LINE-INSERTION-PROCESS-FAILED, with a start time of 13.53 (first insertion event) and an end time of 14.09 (last removal event). BT-45 expresses this as *There were 3 failed attempts to insert a peripheral venous line at 13:53*; the time given is the time the abstract event started, which is reasonable in this case. Now, if the final insertion attempt at 14.08 had been successful, the BT-45 data abstraction module would have instead produced the abstract event LINE-INSERTION-PROCESS-SUCCEEDED, with similar times, and BT-45 would have produced the text

> After three attempts, at 13.53 a peripheral venous line was inserted successfully.

In other words, the time given would still be the time that the abstract event started; but this is misleading, because readers of the above text expect that the stated time is the time of the successful insertion (14.08), not the time at which the sequence of insert/remove events started.

We need a much better model of how to communicate time, and how this communication depends on the semantics and linguistic expression of the events being described. An obvious first step, which we are currently working on, is to include a linguistically-motivated temporal ontology (Moens and Steedman, 1988), which will be separate from the existing domain ontology. We also need better techniques for communicating the temporal relationships between events in cases where they are not listed in chronological order (Oberlander and Lascarides, 1992).

6 Discussion

Two discourse analysts from Edinburgh University, Dr. Andy McKinlay and Dr Chris McVittie, kindly examined and compared some of the human and BT-45 texts. Their top-level comment was that the human texts had much better narrative structures than the BT-45 texts. They use the term 'narrative' in

the sense of Labov (1972, Chapter 9); that is story-like structures which describe real experiences, and which go beyond just describing the events and include information that helps listeners make sense of what happened, such as abstracts, evaluatives, correlatives, and explicatives.

Dr McKinlay and Dr. McVittie pointed out many of the problems mentioned above, but they also pointed out a number of other narrative deficiencies in the BT-45 texts. The most fundamental was that the human texts did a much better job of linking related events into a coherent whole. Other deficiencies include the lack of any kind of conclusion in the BT-45 texts.

We agree with this analysis; it is striking that many of the specific problems identified are related to the problem of generating narratives. Continuity, description of related channels, overview of behaviour over time, and communication of time are all aspects of narrative in the broad sense; they are things we need to get right in order to turn a text into a story. This point is especially significant in light of the fact that many of our medical collaborators at Edinburgh have informally told us that they believe stories are valuable when presenting information about the babies, and indeed that a major problem with data visualisation systems compared to written notes (which they used many years ago) is that the visualisation systems do not tell stories.

Unfortunately, we are not aware of any previous research in the NLG community about these issues. Researchers in the creativity community have looked at issues such as plot and character development in systems that generate fictional stories (Perez y Perez and Sharples, 2004); but this is not relevant to our problem, which is presenting non-fictional events as a narrative. Callaway and Lester (2002) looked at microplanning issues in narrative generation, including reference, lexical variation, and aggregation; but none of these were identified in our evaluation as major problems in text quality.

7 Future Work

The BABYTALK project continues until August 2010, and during this period we hope to investigate most of the issues identified above, especially the ones related to narrative. We are currently conducting experiments to improve the way we communicate time, and we have started redoing the document planner to do a better job of describing systems of related channels in a unified manner. We are also investigating top-down data abstraction and document planning approaches which we hope will address continuity problems, and which may assist in better overviews and narrative structures. We are also working on many issues not directly related to narrative, such as reasoning about and communicating uncertainty, use of vague language, generation of texts for non-specialists (e.g., parents), and HCI issues.

We would welcome interest by other researchers in these topics (there is more that needs investigating than we can do on our own!), and we would be happy to assist such people, for example by sharing some of our code and data resources.

8 Conclusion

We believe that there is enormous potential in systems such as BABYTALK which generate textual summaries of data; the world desperately needs better techniques to help people understand data sets, and our experiments suggest that good textual summaries really can help communicate data sets, at least in some contexts. However, building good data summarisation systems requires the NLG research community to address a number of problems which it has not traditionally focused on, many of which have to do with generating good narratives. We intend to focus much of our energy on these issues, and would welcome research contributions from other members of the community.

Acknowledgements

Many thanks to our colleagues in the BabyTalk project, and to the doctors and nurses who participated in the evaluation; this work would not have been possible without them. Special thanks to Dr McKinlay and Dr McVittie for agreeing to analyse the texts for us. We are also grateful to our colleagues in the Aberdeen NLG group, and to the anonymous reviewers, for their helpful comments. This research was funded by the UK Engineering and Physical Sciences Research Council, under grant EP/D049520/1.

References

A Belz and E Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proceedings of EACL-2006*, pages 313–320.

C Callaway and J Lester. 2002. Narrative prose generation. *Artificial Intelligence*, 139:213–252.

G Carenini and J Moore. 2006. Generating and evaluating evaluative arguments. *Artificial Intelligence*, 170:925–952.

P Carpenter and P Shah. 1998. A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, 4:74–100.

L Ferres, A Parush, S Roberts, and G Lindgaard. 2006. Helping people with visual impairments gain access to graphical information through natural language: The iGraph system. In *Proceedings of ICCHP-2008*.

E Goldberg, N Driedger, and R Kittredge. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53.

W Labov. 1972. *Language in the Inner City*. University of Pennsylvania Press.

A Law, Y Freer, J Hunter, R Logie, N McIntosh, and 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.

M Moens and M Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

J Oberlander and A Lascarides. 1992. Preventing false temporal implicatures: Interactive defaults for text generation. In *Proceedings of COLING-1992*, pages 721–727.

R Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of LREC-2006*.

R Perez y Perez and M Sharples. 2004. Three computer-based models of storytelling: Brutus, Minstrel, and Mexica. *Knowledge-Based Systems*, 17:15–29.

F Portet, E Reiter, J Hunter, and S Sripada. 2007. Automatic generation of textual summaries from neonatal intensive care data. In *Proceedings of AIME 2007*.

R Power, D Scott, and N Bouayad-Agha. 2003. Document structure. *Computational Linguistics*, 29:211–260.

E Reiter, S Sripada, J Hunter, J Yu, and I Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167:137–169.

E Reiter. 2007. An architecture for Data-to-Text systems. In *Proceedings of ENLG-07*, pages 97–104.

S Sripada, E Reiter, and L Hawizy. 2005. Evaluation of an NLG system using post-edit data: Lessons learned. In *Proceedings of ENLG-2005*, pages 133–139.

R Turner, S Sripada, E Reiter, and I Davy. 2008. Using spatial reference frames to generate grounded textual summaries of georeferenced data. In *Proceedings of INLG-2008*.

M van der Meulen, R Logie, Y Freer, C Sykes, N McIntosh, and J Hunter. submitted. When a graph is poorer than 100 words: A comparison of computerised natural language generation, human generated descriptions and graphical displays in neonatal intensive care.

J Yu, E Reiter, J Hunter, and C Mellish. 2007. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13:25–49.