Automatic Generation of Textual Summaries from Neonatal Intensive Care Data

François Portet, Ehud Reiter, Jim Hunter and Somayajulu Sripada

Department of Computing Science, University of Aberdeen, Aberdeen AB24 3UE, UK
{fportet,ereiter,jhunter,ssripada}@csd.abdn.ac.uk

Intensive care is becoming increasingly complex. If mistakes are to be avoided, there is a need for the large amount of clinical data to be presented effectively to the medical staff. Although the most common approach is to present the data graphically, it has been shown that textual summarisation can lead to improved decision making. As the first step in the BabyTalk project, a prototype is being developed which will generate a textual summary of 45 minutes of continuous physiological signals and discrete events (e.g.: equipment settings and drug administration). Its architecture brings together techniques from the different areas of signal analysis, medical reasoning, and natural language generation. Although the current system is still being improved, it is powerful enough to generate meaningful texts containing the most relevant information. This prototype will be extended to summarize several hours of data and to include clinical interpretation.

1 Introduction

In the Intensive Care Unit (ICU), the interpretation of clinical data is an essential task. Generally, the data available consists of (i) continuously monitored physiological variables (e.g.: heart rate, blood pressure, etc.) and (ii) discrete events (e.g.: equipment settings, drug administration, etc.). However the volumes of data are so large (about 1 MB per patient per day), that attention overload (looking after several patients) and stress can lead to mistakes being be made. Hence, presenting these data in an efficient way is crucial for informed medical decision making.

Although the graphical presentation of data is becoming standard practice, an offward experiment conducted as part of the Neonate project (Hunter *et al.*, 2003) showed that medical professionals, in some circumstances, are more likely to make better treatment decisions if they are given a textual summary of patient data, instead of a graphical one (Law *et al.*, 2005). In this *GraphVsText* experiment, forty nurses and doctors with different levels of expertise were asked (individually) to say what actions(s) they would take for a baby whose recent history over a period of about 45 minutes was presented either graphically (as time series plots) or in the form of a textual summary; the text was written by senior clinicians based on the graphical presentation. Each participant was presented with 16 cases, eight graphical and eight textual. Although the clinicians said they preferred the more familiar graphical presentation, they chose more correct actions after reading the textual summaries. Recent experi-

mental research comparing textual and graphical summaries of mobile phone manuals (Lagan-Fox *et al.*, 2006) also showed a decision-making improvement with texts.

These results motivated the BabyTalk[1] project whose goal is the automatic generation of texts summarising baby's ICU data in neonatal units. Over the past decade, we have acquired considerable experience in Data Analysis in neonatology with the Cognate (Logie *et al.*, 1997) and Neonate projects as well as in the generation of text from discrete and continuous data with the SunTime (Sripada *et al.*, 2002, Yu *et al.* 2007) project. A number of techniques have been developed elsewhere for summarising clinical data. For example, the CLEF project (Hallett *et al.*, 2005) aims at generating summaries of multiple text-based health reports. Perhaps the most successful applications have been tools that (partially) automate the process of writing routine documents, such as Hüske-Kraus's Suregen system (Hüske-Kraus, 2003a), which is regularly used by physicians to create surgical reports; see Hüske-Kraus, 2003b for a review of text generation in medicine. However, the complete summarisation of ICU data is more complex, involving the processing of time series, discrete events, and short free texts, which seems to have not been done before.

The BabyTalk project aims at providing summaries according to two different dimensions: duration and degree of abstraction. Four mains systems are planned:

1. **BT-45:** descriptive summary of 45 minutes data
2. BT-Nurse: summary of 12 hours of data to serve as a shift summary.
3. BT-Doc: similar to BT-nurse but with the intention of supporting decision making by junior doctors.
4. BT-Family: a daily summary for the baby's family, adapted to the emotional state of the recipient.

BT-45 is the simplest system as it covers a limit time period and is purely descriptive. It will be used as a stepping stone to the development of the other systems which involve longer durations and more interpretation. BT-45 will be evaluated by repeating the *GraphVsText* experiment with *three* types of presentation: graphical, text written by experts, text generated automatically. This paper describes the progress we have made towards the implementation of BT-45. The data that BT-45 must deal with and the target textual outputs are presented Section 2. The architecture is discussed in Section 3. The prototype has been informally tested by comparison with the manually generated summaries and the results of this comparison are presented in Section 4. The paper ends with a discussion about necessary improvements and other future activities within the project.

Although our work is carried out in the context of neonatal intensive care, we expect the principles to be applicable more widely to adult ICU and to other high-dependency units.

[1] http://www.csd.abdn.ac.uk/research/babytalk/

2 Inputs and Outputs

2.1 Input data

The inputs to BT-45 are of two kinds: (i) continuous multi-channel time series data from the physiological monitors and (ii) discrete event data such as the entry of laboratory results, actions taken, etc.

Physiological time series data
A maximum of seven channels were recorded: the Heart Rate (HR), the pressures of oxygen and carbon dioxide in the blood (OX and CO), the oxygen saturation (SO), the peripheral and central temperatures of the baby (TP and TC) and the mean blood pressure (BM). We have over 400 hours of continuously recorded data from babies in the Neonatal Intensive Care Unit at the Edinburgh Royal Infirmary. As with all real ICU data, our data are sometimes incomplete (periods for which some probes are off) and contain periods of noise.

Discrete data
As part of the Neonate project (Hunter *et al.*, 2003) we employed a research nurse to be present at the cot-side and to record *all* of the following types of event:

- the **equipment** used to monitor, ventilate, etc.;
- the **settings** on the various items of equipment (including the ventilator);
- the results of **blood gas** analysis and other **laboratory results**;
- the current **alarm limits** on the monitors;
- the **drugs** administered;
- the **actions** taken by the medical staff;
- occasional descriptions of the physical state of the baby (**observations**);

The unit we are working with is about to go 'paperless'. This means that we can expect that the equipment used, settings, lab and blood gas results, alarm limits and medication will be automatically recorded. However, human activities such as actions and observations are more difficult to acquire electronically and it is not clear at present exactly what will be recorded and with what timing accuracy.

For the implementation of BT-45 we will use the data collected by the research nurse. However we realise that for systems to operate in the real world, they will only be able to access that information which is available to them electronically and it is part of our research agenda so see to what extent missing items can be derived from what *is* available.

2.2 Output data

The summaries against which we will compare our automatically generated texts were created by a consultant neonatologist and an experienced neonatal nurse researcher. In order to obtain a comparison with the graphical presentation which was as valid as possible, we took steps to ensure that the texts were purely descriptive (i.e. did not

contain higher level clinical interpretations) and contained information that came from the 45 minute data period only; 18 summaries were generated for time-periods varying between 30 minutes and 53 minutes (mean = 40.5). An example appears in Fig 5.

3 Architecture

The main architecture of the prototype is shown in Fig. 1. BT-45 creates a summary of the clinical data in four main stages. The physiological time series and the annotations are processed by **Signal Analysis (1)** to abstract the main features of the signals (artifacts, patterns, and trends). **Data Abstraction (2)** performs some basic reasoning to infer relations between events (i.e.: "A" causes "B"). From the large number of propositions generated, **Content Determination (3)** selects the most important**,** and aggregates them into a tree of linked events. Finally, **Micro Planning and Realization (4)** translates this tree into text. All of the terms used to describe the discrete events are **described by** an **Ontology (5)** of NICU concepts. These were mainly acquired during the Neonate project and are still being extended.

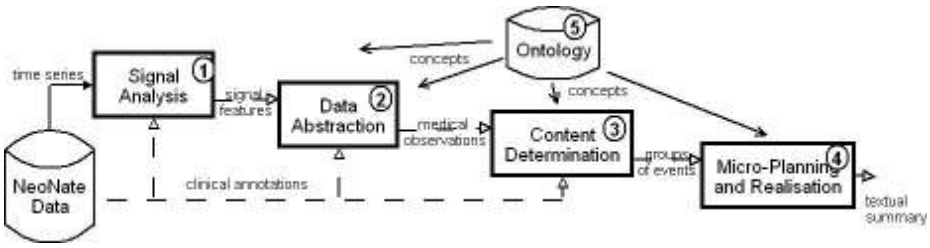


Fig. 1. Architecture of BT-45.

3.1 Signal Analysis

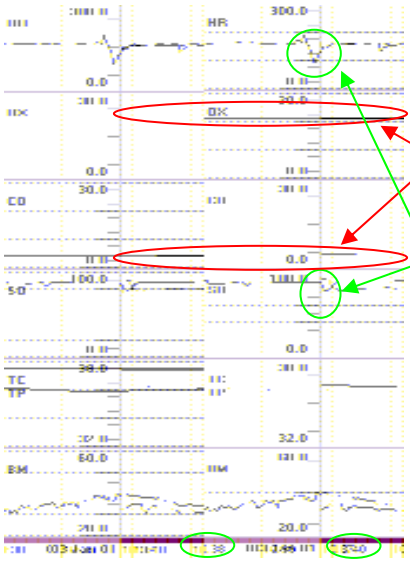
This module analyses the time series data to detect artifacts, patterns, and trends.

An artifact is defined as a sequence of signal sample values that do not reflect real physiological data. In BT-45, the first stage of detection consists of simple thresholding of the impossible values (mainly due to a probe falling off, being partially detached, being removed by a nurse, etc.). For example a heart rate cannot be 0 and a baby temperature cannot be physiologically below 30 degrees Celsius. The artifact time intervals are then merged if they are close to each other (within 10 sec). The second stage of the artifact detection is performed by an expert system which relates the artifacts between the different channels. For example, as the OX and CO channels are derived from the same probe (the transcutaneous probe), if an artifact appears on one channel, it should also appear on the other.

Pattern recognition is based on the rapid-change detector of the *SumTime-Turbine* project (Yu *et al.* 2007) and looks for cases where the signal data is changing rapidly. Pattern intervals are created by merging nearby rapid-change points, and these are then classified into two kinds of patterns, spike and step, using heuristics.

Trend detection uses bottom-up segmentation (Keogh *et al.*, 2001). The code is a simplified version of the segmentation of the SumTime-Mousam project (Sripada *et al.*, 2002). Bottom-up segmentation consists in merging neighbouring segments iteratively into larger ones. Before the merging, two neighbourhood segments are approximated by a line and if the error is less than a specific threshold then the segments are merged. The operation is repeated until the total error reaches a specific threshold or only one segment remains. In this implementation, every sample of the time series belonging to an artifact or a spike is ignored. This enables us to acquire the longer-term trends of a time series rather than rapidly changing features.

The output of Signal Analysis consists of events with a stated duration. For example, for scenario 1 during the period 10:38 to 10:40, the events presented Fig. 2 are generated. Each line consists of: **event type** (**channel**), **start time**, **end time** (**importance**)", where importance is scored from 0 to 100. The first line shows that samples of the OX channel have been classified as artefact and the main shape of these samples corresponds to a downward spike. As OX and CO come from the same probe, the second stage of the artefact detection inferred the same period as artifact on the CO channel. Two rapid changes have been detected by the pattern recognizer on the HR and SO channel. Then trends have been established for other channels. Note that the computation of the upward trend on HR did not take into account the period during which a downward spike was detected.



Artefacts:

ARTEFACT;SPIKE;DOWN (OX):	10:30:00	11:11:59 (0)
ARTEFACT;SPIKE;DOWN (CO):	10:30:00	11:11:59 (0)

Patterns:

PATTERN;SPIKE;DOWN (HR):	10:39:22	10:40:06 (14)
PATTERN;STEP;DOWN (SO):	10:39:39	10:40:05 (3)

Trends:

TREND;UP (HR):	10:30:00	10:49:52 (5)
TREND;FLAT (TC):	10:30:00	10:46:44 (0)
TREND;FLAT (TP):	10:30:00	10:46:17 (0)
TREND;UP (BM):	10:30:00	10:47:24 (4)
TREND;DOWN (SO):	10:30:00	10:40:10 (11)

Fig. 2. Input and Output of Signal Analysis for scenario 1 from 10:38 to 10:40.

3.2 Data Abstraction

Data abstraction consists in finding pairs of events that are related by heuristics. There are three kinds of link: **causes**, **includes** and **associates**. For examples, if a bradycardia is found during an intubation then this intubation is the likely **cause** of the bradycardia; **includes** is used for events that are always accompanied by other events (e.g.: hand-bagging is included in intubation), and **associates** is for obvious correlations (e.g.: overlapping spikes in OX and CO are associated as they come from the same probe).

The output of the Data Abstraction module consists of relations between events. Fig 3 shows the results of the processing of the events of Fig. 2. The first link found associates a downward step on SO with a downward spike on HR. The rule stated that if there are two downward patterns on HR and SO during a short period then they should be associated with the same external phenomenon. The last link states that the decrease in SO should have caused the increase of the FiO2 (fraction of inspired oxygen). Indeed, the SO is known to be influenced by the nurse increasing the FiO2 in cases of desaturation.

```
Link (ASSOCIATED)
   PATTERN;STEP;DOWN (SO):                 10:39:39    10:40:05 (3)
   PATTERN;SPIKE;DOWN (HR):                10:39:22    10:40:06 (14)
Link (ASSOCIATED)
   PATTERN;STEP;DOWN (SO):                 10:39:39    10:40:05 (3)
   PATTERN;SPIKE;DOWN (HR):                10:40:27    10:41:31 (4)
Link (ASSOCIATED)
   PATTERN;STEP;DOWN (SO):                 10:39:39    10:40:05 (3)
   TREND;UP (SO):                          10:40:10    10:47:56 (10)
Link (CAUSES)
   SETTING;VENTILATOR;FiO2 (35.0):         10:38:38    10:38:38 (2)
  TREND;UP;SO (SO):                        10:40:10    10:47:56 (10)
Link (CAUSES)
   TREND;DOWN (SO):                        10:30:00    10:40:10 (11)
   SETTING;VENTILATOR;FiO2 (36.0):         10:30:10    10:30:10 (10)
```

Fig. 3. Output of Data Abstraction.

3.3 Content Determination

Content Determination decides what information needs to be communicated in the text, and how this information should be structured. To do so, events are grouped according to relational links between them in a way similar to Hallet *et al.* (2005). The event groups are then used to compose the tree for the document. The decision as to which groups to mention in the text is based on heuristics which try to produce a document of a certain length. A pruning of events is also based on their importance and on a notion of not mentioning events which the reader will infer by herself. Content Determination also generates introductory information which describes the state of the baby at the start of the scenario. Finally, the components are temporally ordered, and linked with Temporal-Sequence relations.

Fig 4 shows the grouping derived from the linked events from Fig 2 and Fig 3. The left side of the figure shows a group of temporally ordered events and the links between events; dashed arrows represent associate links and full arrows causal links. The right side of the figure shows the tree composed from the group; the number of

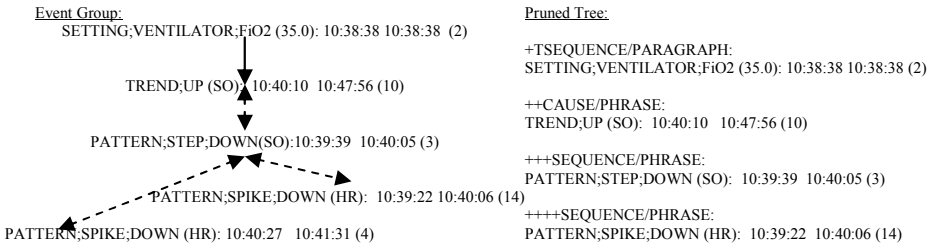


Fig. 4. Group of liked events and the corresponding pruned tree.

"+" signs gives the depth of the node) which is composed of phrases of different types. The reference to FiO2 forms the beginning of the paragraph and it is followed by a causal relation phrase and two other phrases. The pruning removes unimportant subtrees; that is why the leaf PATTERN;SPIKE;DOWN (HR): (4) has been remove but not the node PATTERN;STEP;DOWN(SO): (3) because it is the parent of the important leaf PATTERN;SPIKE;DOWN (HR): (14).

3.4 Microplanning and Realisation

The Microplanner converts the tree into text. Most events are converted into simple syntactic structures using mapping rules. For example, increasing the ventilator FiO2 level is converted into a phrase with verb "increase", object "FiO2", a prepositional phrase saying "to XX" (where XX is the new level), and possibly another prepositional phrase saying "at TT" (where TT is the time of the action). The tree (events related by discourse relations) is linearised top-down. A parent is always expressed before its children. The order that the children are expressed in depends on their discourse relation to their parent.

The realisation performs the actual translation into text from the output of the microplanner via a syntactic structure and completes the morphology and the layout of the document. The realisation of the tree of the Fig 4 gives:

"At 10:38 FiO2 is increased to 35 **so** saturations rise for 8 minutes to 96 **and** saturations drop to 78 **and** there is a bradycardia to 90".

Each event is converted by microplanning. The causal phrase is linked to the previous one by "**so**" and the sequence phrases are linked by an "**and**" to form a single sentence.

4 Preliminary results

BT-45 is currently being developed and will be evaluated experimentally by comparing graphical presentation, human textual summaries and computerized textual summaries following the method described in Law *et al.* (2005). We have presented some outputs to a senior neonatologist who was very impressed by the generated texts even

if they are not exploitable at present. As a general criterion, the length of the texts has been computed. This showed that our outputs were generally too long (median of 138 words) and with a greater standard deviation (56.4 words) than the human summaries (median of 110 words with STD = 41.1 words).

To demonstrate the performance of the prototype, we present an informal comparison of computerized and expert summaries for scenario 1, represented in Fig. 5. Scenario 1 concerns the re-intubation of a baby. The action we expected to be recommended at the end of the period was the decision to X-ray the baby in order to check the position of the endo-tracheal tube. Consequently, information about intubation, suction and the physiological variables are important.

BT-45

*** introduction ***

At 10:30 you see the baby.

HR = 148, mean BP = 28, central temperature = 37.5, peripheral temperature = 36.3 and sats = 96.

*** 1 ***

Saturations fall for 10 minutes to 77 so FiO2 is decreased to 36.

At 10:38 FiO2 is increased to 35 so saturations rise for 8 minutes to 96 and saturations drop to 78 and there is a bradycardia to 90.

*** 2 ***

At 10:46 peripheral temperature falls for 13 minutes to 33.5 and then central temperature falls for 5 minutes to 36.5.

At 10:51 the baby is intubated and so there is a significant bradycardia to 61, there is a desaturation to 77 and mean BP jumps to 47. As part of this procedure, at 10:47 the baby is hand-bagged so there is a bradycardia to 125.

At 10:52 toe/core temperature gap rises for 7 minutes to 2.4, at 11:02 FiO2 is decreased to 67 and FiO2 is increased to 79.

*** 3 ***

Saturations drop to 79 and then at 11:05 saturations rise for 6 minutes to 99 so FiO2 is decreased to 80.

The baby is sucked out and at 11:09 FiO2 is increased to 61.

EXPERT

*** introduction ***

You see the infant first at 1030.

The transcutaneous OX/CO electrode is being recalibrated.

*** 1 ***

In preparation for re-intubation, a bolus of 50ug of morphine is given at 1039 when the FiO2 = 35%. There is a momentary bradycardia and then the mean BP increases to 40. The sats go down to 79 and take 2 mins to come back up. The toe/core temperature gap increases to 1.6 degrees.

*** 2 ***

At 1046 the baby is turned for re-intubation and re-intubation is complete by 1100 the baby being bagged with 60% oxygen between tubes. During the re-intubation there have been some significant bradycardias down to 60/min, but the sats have remained OK. The mean BP has varied between 23 and 56, but has now settled at 30. The central temperature has fallen to 36.1°C and the peripheral temperature to 33.7°C. The baby has needed up to 80% oxygen to keep the sats up.

*** 3 ***

Over the next 10 mins the HR decreases to 140 and the mean BP = 30-40. The sats fall with ETT suction so the FiO2 is increased to 80% but by 1112 the FiO2 is down to 49%.

Fig. 5. Examples of BT-45 and expert summaries for the same scenario

- BT-45 introduced the start time and values of the physiological variables (as is normally the case) whereas, this time, the human writer didn't. Dealing with the multiplicity of styles in human writers is an important problem out of the scope of this paper (see Reiter *et al.*, 2005). Moreover, BT-45 did not indicate that OX and CO are being re-sited (hence the values are wrong), but, at least the OX and CO channels have been detected as containing artifact and not included in the text.

- The first part of the summary addresses the correspondence between desaturation and the FiO2 settings. This information is also present in the human summary but in a more condensed way and with more information about medication and intubation. This information is available to BT-45 and rules about the protocols for intubation should be added to the Data Abstraction module.
- The second part concerns the re-intubation. In this important period, BT-45 succeeded in tying the bradycardia and hand-bagging to the intubation. The temperature, even if not summarised enough is also described. The desaturation event seems to contradict the human text: "but the sats [SO] have remained OK". This is due to the expert's view that this desaturation (actually present in the data) is not relevant as far as the intubation is concerned.
- The third part is about the saturation problems following the intubation. BT-45 detected the falls in saturation and related them to the FiO2 setting and the suction but didn't mention the fall in heart rate and the variation in mean BP.

This informal comparison enabled us to identify three main problems:

- Crucial information about medication and medical activities such as intubation must be handled by the system by increasing significantly the number of expert rules in the Data Abstraction module.
- The lack of aggregation (e.g.: the FiO2 is increased, and the FiO2 is decreased, etc.) leads to texts which are too long and too far from the human style. The aggregation could be performed in the Data Abstraction module by a mechanism that groups the overlapping events of same type into a "sequence of events".
- Information is not highlighted in the text. Important event must be emphasized and less important events must be hidden. This is highly dependent on context e.g. if the baby is being intubated or is under specific medication.

Despite these drawbacks (which are being addressed in the next version) the BT-45 output contains the most important information in this scenario: intubation, hand bagging, suction, desaturation and bradycardia.

5 Discussion

BT-45 is an ongoing project and there is much to do in order to reach the quality of the experts' text (if indeed this is possible). However, our prototype has demonstrated that it is possible to perform simple data analysis and reasoning that are sufficient to generate a text where the most important information is presented.

Although the Signal Analysis module is composed of simple algorithms which perform in satisfactory way in our scenarios, they must be extended to deal with noisier data. Many artifact removal algorithms exist and an experiment to compare a number of them on noisy neonatal data is planned. Another issue is the detection of human activities from the signal (e.g.: re-siting of probes). For this, an approach based on trend and syntactic analysis will be investigated (Hunter and McIntosh, 1999).

The Data Abstraction is for the moment very basic. At this level we need to generate more high level abstractions such as the qualification of events (e.g.: "significant"

bradycardia), aggregation of similar terms, more linking, etc. This is the weakest component of BT-45 at present and improvements will be based on the acquisition and formalisation of a greater range of expert knowledge.

The way the Content Determination module selects and organizes the information to present in the text is mainly based on the importance factor of the events. This clearly needs to be sensitive to the protocols/procedures being applied and to the characteristics of different babies. Thus, the decision as to which information to hide and to show is an important issue.

The microplanning and realization translates the tree of events into text. It should be more sensitive to aggregation and reference that can be controlled by stylistic parameters (such as desired sentence length). Then, a general lexicalisation engine must be set up to control the usage of technical vocabulary and vague modifiers (such as "small" spike). Moreover, a technique must be implemented to control the multiple time references in the text.

References

Hallett C and D Scott. 'Structural variation in generated health reports', Proceedings of the 3rd International Workshop on Paraphrasing, Jeju Island, Korea (2005).

Hunter JRW and N McIntosh. 'Knowledge-Based Event Detection in Complex Time Series Data'. AIMDM 1999, Aalborg, pp 271-280 (1999).

Hunter JRW, L Ferguson, Y Freer, G Ewing, R Logie, P McCue and N McIntosh. 'The NEONATE Database', Workshop on Intelligent Data Analysis in Medicine and Pharmacology and Knowledge-Based Information Management in Anaesthesia and Intensive Care, AIME-03, pp 21-24 (2003).

Hüske-Kraus D. 'Suregen-2: A Shell System for the Generation of Clinical Documents', Proceedings of EACL-2003 (demo session) (2003a).

Hüske-Kraus D. 'Text Generation in Clinical Medicine – a Review', Methods of Information in Medicine, 42, pp 51-60 (2003b).

Keogh E, S Chu, D Hart and M Pazzani. 'An Online Algorithm for Segmenting Time Series', Proceedings of IEEE International Conference on Data Mining. pp 289-296 (2001).

Langan-Fox J, C Platania-Phung and J Waycott. 'Effects of Advance Organizers, Mental Models and Abilities on Task and Recall Performance Using a Mobile Phone Network', Applied cognitive psychology, 20, pp 1143-1165 (2006).

Law AS, Y Freer, JRW Hunter, RH Logie, N McIntosh and J Quinn. '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, pp 183-194 (2005).

Logie RH, JRW Hunter, N McIntosh, KJ Gilhooly, E Alberdi and J Reiss. 'Medical Cognition and Computer Support in the Intensive Care Unit: A Cognitive Engineering Approach', Engineering Psychology and Cognitive Ergonomics: Integration of Theory and Application, pp 167-174 (1997).

Reiter E, S Sripada, J Hunter, J Yu and I Davy. 'Choosing Words in Computer-Generated Weather Forecasts', Artificial Intelligence, 167, pp 137-169 (2005).

Sripada S G, E Reiter, J Hunter and J Yu. 'Segmenting Time Series for Weather Forecasting', Applications and Innovations in Intelligent Systems X, pp 193-206 (2002).

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