Artificial Intelligence in Medicine xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse

James Hunter[a,*], Yvonne Freer[b], Albert Gatt[a,c], Ehud Reiter[a], Somayajulu Sripada[a], Cindy Sykes[b]

[a] Department of Computing Science, University of Aberdeen, King's College, Aberdeen AB24 3UE, UK
[b] Simpson Centre for Reproductive Health, Royal Infirmary of Edinburgh, EH16 4SA, UK
[c] Institute of Linguistics, University of Malta, Msida MSD 2080, Malta

ARTICLE INFO

Article history:
Received 15 March 2012
Received in revised form
11 September 2012
Accepted 19 September 2012

Keywords:
Natural language generation
Natural language processing
Data to text
Neonatal intensive care
Health informatics

ABSTRACT

Introduction: Our objective was to determine whether and how a computer system could automatically generate helpful natural language nursing shift summaries solely from an electronic patient record system, in a neonatal intensive care unit (NICU).
Methods: A system was developed which automatically generates partial NICU shift summaries (for the respiratory and cardiovascular systems), using data-to-text technology. It was evaluated for 2 months in the NICU at the Royal Infirmary of Edinburgh, under supervision.
Results: In an on-ward evaluation, a substantial majority of the summaries was found by outgoing and incoming nurses to be understandable (90%), and a majority was found to be accurate (70%), and helpful (59%). The evaluation also served to identify some outstanding issues, especially with regard to extra content the nurses wanted to see in the computer-generated summaries.
Conclusions: It is technically possible automatically to generate limited natural language NICU shift summaries from an electronic patient record. However, it proved difficult to handle electronic data that was intended primarily for display to the medical staff, and considerable engineering effort would be required to create a deployable system from our proof-of-concept software.

1. Introduction

In the 21st century, increasing volumes of information about patients are available to medical practitioners, particularly in the intensive care unit (ICU), where the data may consist of both continuously monitored physiological data (e.g. heart rate) and detailed records of interventions and observations. The availability of these data raises the expectation that clinical professionals will make best use of them to improve patient care, avoid errors and keep patients informed. If this expectation is to be realised, the recipients must not become overwhelmed by the volume of the data; information must be presented in a way that can be easily assimilated and that emphasises what is medically important.

For decision-making in real time it is clearly important that the data be presented effectively so that they can be easily understood. However appropriate presentation is equally important when a patient is handed over from one clinician to another. The incoming clinician needs to understand the patient's history, problems, response to recent interventions, etc. In most cases there is an oral or written handover between outgoing and incoming clinicians, but in practice the outgoing clinician may forget to mention important information, or the incoming clinician may find it difficult to assimilate everything presented to her during a brief exchange. In such cases the incoming clinician relies on the available data in order to fill in gaps in her understanding of the patient. However, this data is likely to be voluminous and heterogeneous, and the process of identifying important elements and assimilating them into a bigger picture about the patient becomes non-trivial.

Currently, computer-generated graphs and/or tables are often used to present large data sets to clinicians. However, some studies have found that visualisations of medical data do not always enhance decision-making [1–3]; other studies have shown that in some circumstances, good-quality textual (in our case, English) summaries of data can be more effective than visualisations as a decision-support tool [4,5]. These studies used summaries written by human experts in a research context, but of course in most circumstances it is impractical for clinicians themselves to spend time writing summaries of data for other clinicians.

We have developed a natural language generation (NLG) system, BT-Nurse [6], which automatically generated English summaries from the electronically recorded patient data over a 12 h nursing shift, for a baby in a neonatal intensive care unit (NICU). BT-Nurse was part of a larger project, BabyTalk [7,8], which was concerned with several types of summary texts which were

* Corresponding author. Tel.: +44 1224 272287; fax: +44 1224 273422.
E-mail address: j.hunter@abdn.ac.uk (J. Hunter).

0933-3657/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2012.09.002

computer-generated from NICU data, including texts for parents and relatives as well as texts for nurses. The first system to be developed within BabyTalk was BT-45 [7], an experimental prototype which summarised 45 min of NICU data (hence the name BT-45) with the intention of supporting decision-making in real time by clinical staff. BT-45's input data included both the continuously monitored physiological data and supplemental discrete information which was collected for the purpose on the ward by a research nurse.

In contrast to BT-45, BT-Nurse summarised 12 h worth of NICU data with the intention of supporting shift handover – the handover from one nurse to another at the end of the nursing shift. Furthermore, BT-Nurse was intentionally designed to use only data that were *routinely* available in electronic form – *no* additional data input and hence *no* additional work from the nurses was required. In summary, the design and implementation of BT-Nurse sought to address the following principal questions:

1. In the light of the volume and heterogeneity of data collected in the NICU, is it feasible to select and abstract the *relevant* data that a nurse would find most helpful?
2. Given appropriate selection and abstraction of such data, can this be automatically summarised in English, using language that is fluent and easily readable?
3. Is the technology underlying such systems robust enough to be used in real scenarios with previously unseen data, collected onward?
4. Is shift summary generation which is based exclusively on electronic data, perceived as understandable, accurate and helpful by its target user population?

These questions were addressed through the design of a proof-of-concept system that was evaluated on-ward for an extended period, with previously unseen data collected live. Although we made extensive use of existing technologies drawn from such fields as artificial intelligence, intelligent signal processing and natural language generation, we made little attempt to make significant advances there. Our hope was rather that a positive reply to our research questions would suggest that these underlying technologies could be integrated and practically deployed in a complex, fault-critical scenario to help medical decision-making. Anticipating the outcomes of the evaluation reported below, our main finding was that a significant majority of our target users found the automatically generated summaries understandable, accurate and helpful. Furthermore, the technology was robust enough to run on live data for an extended period without any serious errors being noted by a senior supervising clinician. Feedback from our users also pointed out some shortcomings, which were primarily due to the system's data processing capabilities in the face of noise and/or information gaps.

Before discussing the design and evaluation of BT-Nurse in more detail, we start by introducing an example. BT-Nurse analysed physiological data for patterns and trends, and integrated this with information about observations and interventions extracted from the electronic patient record. It decided which information was most important and generated an English text which presented this information. An extract from a BT-Nurse text is shown in Fig. 1. Fig. 2 shows the corresponding extract from a summary of the *same* shift which was written by an expert nurse, specifically for the purposes of the BabyTalk project. This shift will be used as a running example throughout this paper.

The remainder of the paper is structured as follows. Section 2 presents relevant background information. In Section 3 we describe how BT-Nurse was designed, implemented and deployed in order to address questions 1 and 2 above. Section 4 presents the results of our evaluation of BT-Nurse carried out on the ward with both outgoing and incoming nurses at the shift handover. This evaluation sought to address our questions 3 and 4. Finally, in Sections 5 and 6 we discuss what we have learned and draw some conclusions.

2. Related work

2.1. Computerised shift handover

Handover is a complex process, and ineffective handover can endanger patient safety; see the review by Friesen et al. [9]. There is no established best practice for handover; most papers on handover are descriptive studies and very few meet medical standards for evidence-based medicine. In their literature review, Friesen et al. stated that the following all contribute to handover problems: failed communication; omitted information; distractions; lack of (or illegible) documentation; lack of utilisation of transfer forms; and lack of easy accessibility to information.

Partly because of the above, there has been an increasing interest in handover systems where part of the handover report is automatically generated by a computer system from the patient record. These systems are based on populating forms and templates using data extracted from an electronic patient record – they do not use NLG technology to summarise the data in the patient record. Menke et al. [10] found that the use of such a system in a paediatric ICU increased the quality and accuracy of documents; however they did not report any significant differences in error rates or patient outcomes. Strople and Ottani [11] made a good case for the desirability of using computers to at least partially automate the shift handover process, but did not describe or evaluate an actual system. However, Staggers et al. [12] found that the computerised patient summary report and the electronic health record were minimally used during the handover and that the existing patient summary reports did not provide adequate cognitive support for nurses. They found that patient summary reports were incomplete, rigid and did not offer "at a glance" information, or help nurses encode information.

Although the beneficial role of automation in clinical settings has been acknowledged, Goddard et al. [13] pointed to some evidence that excessive reliance on automation introduces the risk of "automation bias", that is, a tendency to ignore possible errors introduced, for example, by clinical decision-support systems (CDSS). They noted that this is a relatively under-researched topic; however their review suggested that the role of CDSS should be viewed as complementing, rather than replacing, expert clinical judgment. This was the position taken in the design of BT-Nurse, where the generated summaries were intended to facilitate rather than replace the other modes of communication used in nurse shift handover.

2.2. Data-to-text

BT-Nurse is an example of a *data-to-text* system – a computer program which automatically generates summaries of data sets in English or another natural language [14,15]. Data-to-text systems use signal analysis, artificial intelligence and NLG techniques. They identify and summarise patterns in the data, determine which information is most useful and relevant to the user, and create a natural language text which presents this information in a readable and understandable form. Data-to-text systems can generate very high quality output; indeed in some cases readers prefer texts produced by a data-to-text system over texts written by a person [16].

Data-to-text systems have been developed in many areas, including weather forecasting [16–20], communicating financial and statistical information [21–23], and engineering [24]. Most previous data-to-text systems generated summaries of relatively

Please cite this article in press as: Hunter J, et al. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.09.002

…

The baby was born at 24 weeks weighing 460 g. He is 2 days old, with corrected gestational age of 24 weeks and 2 days, and in an intensive care nursery.

…

Respiratory Support

Current Status

Currently, the baby is on CMV in 27 % O2. Vent RR is 55 breaths per minute. Pressures are 20/4 cms H2O. Tidal volume is 1.5.

SaO2 is variable within the acceptable range and there have been some desaturations.

The most recent blood gas was taken at around 07:45. Parameters are acceptable. pH is 7.3. CO2 is 5.72 kPa. BE is -4.6 mmol/L. The last ET suction was done at about 05:15.

Events During the Shift

A blood gas was taken at around 19:45. Parameters were acceptable. pH was 7.18. CO2 was 7.71 kPa. BE was -4.8 mmol/L.

Another ABG was taken at around 23:00. Blood gas parameters had deteriorated to respiratory acidosis by around 23:00. pH was 7.18. CO2 had risen to 9.27 kPa by around 23:00. BE was -4.8 mmol/L.

The baby was intubated at 00:15 and was on CMV. Vent RR was 50 breaths per minute. Pressures were 20/4 cms H2O. FiO2 was 29 %. Tidal volume was 1.5. He was given morphine and suxamethonium. MAP was raised from 6 cms H2O to 8 cms H2O.

Between 00:30 and 03:15, SaO2 increased from 88 % to 97 %.

Another ABG was taken at around 00:45. pH was 7.18. CO2 dropped to 7.95 kPa. BE was -4.8 mmol/L.

Another blood gas was taken at about 06:15. Blood gas parameters had deteriorated to respiratory acidosis by about 06:15. pH was 7.18. CO2 was 8.4 kPa. BE was -4.8 mmol/L.

Potential Problems

- Purulent secretions during shift suggest risk of infection.

Cardiovascular

…

Fig. 1. Extract from a BT-Nurse summary.

small amounts of homogenous data. In contrast, the clinical record which was the input to BT-Nurse is both large and heterogeneous, containing a mixture of regularly sampled numeric time-series sensor data, laboratory results recorded at irregular intervals, and diverse information entered by doctors and nurses. Handling large heterogeneous data sets is a challenge for data-to-text systems, especially in data analysis and summarisation, and in document planning.

A number of medical informatics systems have been developed which use NLG technology; see the review by Hueske-Kraus [25]. Most of these are not data-to-text systems in the sense of automatically generating linguistic summaries of large clinical data sets; instead they focus on generating patient information material (often to support behaviour change) [26], helping clinicians write routine documents [27], summarising medical literature [28], and generating explanations for recommendations [29].

Two medical systems which summarised information extracted from an electronic patient record were MAGIC [30] and Narrative Engine [31]. MAGIC generated very short summaries of data, which were intended to support post-operative care immediately after surgery (coronary arterial bypass graft); these summaries did not include time-series sensor data. Narrative Engine was a commercial product which generated summaries of a doctor/patient encounter, in part for legal reasons (i.e. to reduce the likelihood of malpractice lawsuits). Since Narrative Engine was a commercial product, its details have not been described in the open literature, but again its summaries do not seem to have included time-series sensor data. Furthermore, Suominen and Salakoski [32] described an approach to extracting textual segments from longer free-text documents, where the selected segments were relevant to a specific topic (e.g. *pain*).

In short, we are not aware of any previous medical data-to-text system which was as ambitious as BT-Nurse (and the other BabyTalk systems described below) in the amount and diversity of data summarised.

2.3. BabyTalk project

The BabyTalk project [7,8] developed a set of data-to-text systems which summarised clinical data from the electronic patient record used in a NICU. Summaries were aimed at a variety of users and purposes. All systems had a similar architecture. In addition to BT-Nurse, the main BabyTalk systems were:

- *BT-45* generated summaries of clinical data over a relatively short period, roughly 45 min, to support decision making by doctors and nurses. BT-45 summaries were typically about half a page

Please cite this article in press as: Hunter J, et al. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.09.002

…
Baby is 24 weeks gestation, weighs 460 grams, and is now 48 hours old.
…

Respiratory support needed due to immaturity

CURRENT MANAGEMENT / ASSESSMENT:

- CMV rate 55, pressures 20/4, in 27% oxygen, giving tidal volumes of 1.5 ml (3.3 ml/kg). Very recent ABG good: pH 7.31, CO_2 5.72. He received morphine prior to intubation at 00:30; no spontaneous respiratory effort noted since being re-ventilated. Desaturates during cares and suction but recovers afterwards; otherwise SpO_2 has been fairly stable. Large ETT secretions, mucopurulent and blood stained.

EVENTS DURING THIS SHIFT:

- While on BiPAP, oxygen requirement increased to 50% by 23:00
- ABG at 23:10 showed CO_2 increased from 7.7 to 9.27 in three hours
- Electively re-intubated at 00:30 to CMV rate 50, pressures 18/4 in 30% oxygen:
  - Difficult intubation; size 2 ETT
  - Was given morphine and suxamethonium
- On ventilation, oxygen requirement reduced to 30% and ABG initially improved
- ABGs worsened; ETT suction yielded large amount secretions; ABG still not improving
- CMV increased to pressures 20/4 and rate 55.

POTENTIAL PROBLEMS and PLAN of CARE:

- Known to have large amount of secretions –
  - Small ETT could become blocked or dislodged
  - assess need for suction regularly
  - check that ETT is secure
- Risk of chest infection related to being ventilated; also due to extreme prematurity and PROM he is at risk of ureaplasma infection – daily ETT secretions samples should be sent for C&S and ureaplasma.

Cardiovascular instability due to immaturity

…

Fig. 2. Extract from a summary written by an expert nurse for the same underlying data as Fig. 1.

long [7]; this system was always intended to be an experimental prototype;
- *BT-Family* generated summaries of 24 h of clinical data, to inform parents of the status of their baby. These summaries were typically about one page long [33].

An additional system, BT-Clan [34], was designed to generate summaries for friends and families (the 'clan'). One reason that BT-Clan was not implemented was that clan members were very interested in the status of the parents as well as the babies, and parent status information was not available in the NICU patient record.

BT-45 was developed first. It was evaluated using a controlled experiment [5], where doctors and nurses were shown clinical information in three presentations (computer-generated text, manually written text, and visualisation), and asked to make a treatment decision; this was done off-ward, using historical data from babies who had been in the NICU several years previously. Decision quality was best when doctors and nurses saw the manually written texts; there was no significant difference in decision quality based on computer-generated texts and visualisations. Analysis of mistaken decisions and comments by doctors and nurses who participated in the experiments highlighted that the computer-generated texts needed to be better narratives [7]; this influenced the design of BT-Nurse.

BT-Family was developed after BT-Nurse, and will be evaluated in the NICU in 2012. It benefitted from lessons learnt in BT-Nurse (Section 4); in particular, far more time was allocated to debugging and testing the system in the NICU.

3. Materials and methods

3.1. The NICU

The NICU provides an environment where pre-term neonates can be cared for while their physiological systems mature, and where sick full-term babies can be treated. As with all ICUs, patients are intensively monitored and nursed, often on a one-to-one basis. The NICU at the Simpson Centre for Reproductive Health at the Edinburgh Royal Infirmary uses the *Badger* system developed by Clevermed[1]® to manage and display patient data. This system acquires and records several channels of continuous physiological data from the cot-side monitor once per second, for example heart rate (HR), oxygen saturation (SaO_2), temperatures (T), and blood pressure (meanBP). These data are presented graphically on a computer located beside each cot. Clinical staff can also use this computer to enter and view additional, discrete items of information, including hourly physiological measurements, drugs and fluids administered, equipment settings, care and treatment actions taken, visual observations, and so on. Most of the data collected are pre-formatted (e.g. using drop-down menus) but free-text entry is also available. The latter is often used when nurses need to comment about a patient's general well-being, as well as a few other observations which are not available within the menus. The system also automatically acquires data when available from laboratory analysers and on-ward blood gas analysers. The NICU is paperless – all relevant data is electronically recorded.

[1] http://www.clevermed.com/.

Please cite this article in press as: Hunter J, et al. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.09.002

07:59 **Nurse Shift Summary**

…

Respiratory

• Respiratory support CMV	
• Inspired oxygen	27.00%
• Oxygen range %	From:27 To:50
• Oxygen saturation range	From:89 To:94
• Respiratory notes	reintubated 00.20hrs size 2 tube at 5.5cm at lips e/t suction x 1 large mucopurulent blood stained secretions. poor gases pressures and rate increased

Cardiovascular

…

Fig. 3. Extract from the *Badger* Nurse Shift Summary for the nursing shift summarised in Fig. 1.

For those observations that are acquired automatically, the time of observation is the same as the time the data is entered. However for the data entered by a user, the *Badger* system does not ask for the time that a particular action was taken or an observation made (the *event time*). These data items are labelled only with the time when they are entered by the user into the database (the *transaction time*), which can occur after some unknown delay relative to the event time. This gave rise to problems which will be discussed later.

One particular display provided by *Badger* is the Nurse Shift Summary which consists of a number of pre-defined data items (including free text entries) collated from all the records entered during a shift. An extract from the Nurse Shift Summary generated by *Badger* for the shift described in Figs. 1 and 2 is shown in Fig. 3. The *Respiratory notes* field here consists of free-text written by the outgoing nurse.

Each nurse in the Simpson NICU cares for one or two babies during a 12 h shift (normally from 08:00 to 20:00 and from 20:00 to 08:00 the following day). At the end of the shift, the handover between nurses takes place on a one-to-one basis at the cot-side. The outgoing nurse will give an oral handover, after which the incoming nurse can look at the *Badger* Nurse Shift Summary and/or the detailed data display. The outgoing nurse does not write her own summary, although she may have completed one or more free text fields which will be included in the *Badger* summary.

3.2. BT-Nurse system architecture

BT-Nurse was constructed around a standard data-to-text 'pipeline' architecture [14,15] where information is[2] processed sequentially by the different modules, as shown in Fig. 4; the modules are activated in the following order:

- data translation;
- signal analysis;
- data pre-processing;
- data interpretation;
- document planning;
- microplanning and realisation.

3.2.1. Ontology

The modules communicate via a domain ontology which defines a standardised terminology; all modules create, inspect, and modify instances in the ontology. We explored the use of a modified version of the UMLS ontology [35] but this proved a poor match to our needs, in particular because we needed to describe observations and actions at a much finer level of detail than was available. We therefore created our own ontology from scratch, which could be integrated with UMLS should the need arise. The BT-Nurse ontology was implemented in Protégé-OWL [36] and contains over 1000 classes. It consists of the following classes and immediate sub-classes. The number after each class shows the number of enclosed sub-classes; examples of further sub-classes are also provided.

- Entity (373)
  - Anatomical abnormality (91); cardiac, chromosome, gastrointestinal, etc.
  - Body system (5): cardiovascular, respiratory, central nervous, gastrointestinal, thermoregulatory.
  - Data channel (96): heart rate, inspired O_2, etc.
  - Instrument (37): arterial line, ET tube, etc.
  - Substance (124): drug, nutrition, etc.
  - Person (20): doctor, nurse, parent, etc.
- Event (650)
  - Action (100): intubation, drug administration, etc.
  - Observation (238): O_2 saturation reading from the monitor, lab result, etc.
  - Phenomenon or process (299): haemetological disorder, infection, etc.
  - Naturally occurring event (13): delivery, death, etc.
- Relation (6 – all sub-classes are given)
  - Association – the weakest type of link where one event is associated in some way with another but the nature of the association is not known or is not important.
    - Contains – a lower level event occurs during a higher level event with which it is associated in some way.
    - Precedes – one event occurs before another.
      - Causes – one event is known physiologically to cause another.
      - Enables – an action is taken so that another action can follow.
      - Triggers – an action is taken in response to an event.

Entities are considered to be objects which do not change over time. *Events*, on the other hand, are temporal objects with a duration (which may be zero) – they can therefore be considered more precisely as intervals [37]. Because (as discussed earlier) the exact start and end times of an event may not be known with certainty, we chose to model this uncertainty by defining the temporal properties of an event by its:

- earliest and latest start times (EST and LST)
- earliest and latest end times (EET and LET)
- minimum and maximum durations (MinD and MaxD)

[2] Although no longer running on the ward, BT-Nurse is still a 'live' system and will be described in the present tense.

Please cite this article in press as: Hunter J, et al. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.09.002

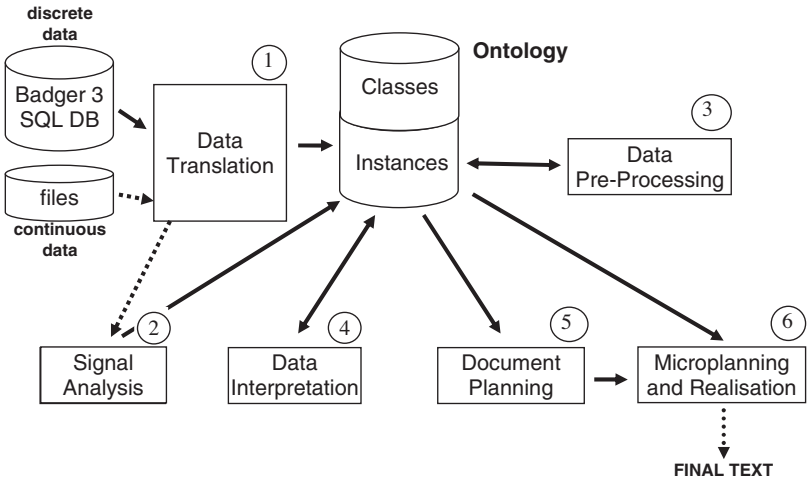

Fig. 4. BT-Nurse architecture: the numbers next to the modules show the order of activation.

where EST is the time before which it is impossible for the event to have started, and so on. For example, we can specify that an intubation (the insertion of a breathing tube down the baby's throat) started at precisely 00:15 (EST = LST = 00:15), or that it started at sometime between 00:00 (EST) and 00:30 (LST). This representation is deliberately identical to that used in the Asbru syntax for representing clinical guidelines [38].

Relations are first-class objects which link two events as set out above; see Section 3.2.6 for examples.

3.2.2. Data translation

The BT-Nurse ontology was, in part, designed to enhance the portability of the BT-Nurse system. The purpose of the data translation module is to read in discrete data from the *Badger* system, and store this information in the ontology. Apart from the obvious conversion from entries in a relational database to instances in the ontology, and the translation of names, a certain amount of restructuring is done during this process. For example, *Badger* uses a single database record to hold information about both an intubation and subsequent extubation (the removal of the tube); the BT-Nurse data translation system splits this record into two and creates separate INTUBATION and EXTUBATION event instances in the ontology. Although most of the translation uses a generalised approach, there are three instances – namely ward/cot locations, drugs, and diagnoses – that require special treatment. In the case of the latter two, this arises because an additional translation from internal *Badger* numerical codes is required.

The data translation system also does a *very* limited amount of free-text information extraction, since some of the information that BT-Nurse needs only occurs in free text comments. This is mainly used for obtaining details of medication. Text is extracted from specific fields in the database which contain nurses' comments on medication. The text is scanned for any tokens that match the names of drugs in the ontology, where partial matches are tolerated within a low edit distance threshold. If the name of a drug is found, the text surrounding the name is further scanned for matches to pre-defined patterns corresponding to drug administration route (e.g. oral), amount (numbers), unit (e.g. mg), action (e.g. start, raise) and type (loading or maintenance dose). The heuristic used here is simply that, given a mention of a drug in the text, the nearest matches containing this additional information must pertain to this drug, unless they are closer to the mention of another drug in the same text. Clearly, these heuristics are far from fully-fledged natural language understanding; indeed, they are limited to known instances of drugs. However, informal evaluation suggested that these heuristics are fairly robust; this is because most text comments in this section of the database usually report one drug administration event and the language used is highly restricted.

Although we believe that we have constructed an ontology which is independent of the detailed structure and coding of the input data, we have only interfaced it to the one dataset (from *Badger*) and this independence has still to be demonstrated in practice, by constructing data translation modules for other datasets.

Another form of data translation is the conversion of the continuous, rapidly sampled data (e.g. heart rate) from the monitor into an internal canonical form. This canonical form is derived from earlier work on signal analysis where it was used with signals from a variety of sources (e.g. gas turbine).

3.2.3. Signal analysis

There are typically several hundred KB of continuous physiological data over a 12-h shift, the exact amount depending on which sensors are deployed on a particular patient. Fig. 5 presents the physiological data corresponding to the summaries in Figs. 1–3. The job of the signal analysis module is to extract a small number (tens or hundreds) of events from this large data set.

For each sensor, six ranges of values are defined:

- *normal*;
- *high*, *low* – unusually high or low but physiologically plausible;
- *very_high*, *very_low* – unusual and of definite medical concern;
- *impossible* – probably arising from noise in the signal.

The boundaries of the *normal* range are derived from the distributions of the historical data from several hundred babies, categorised by gestation and age. The boundaries of the other regions were defined by our experts.

Signal analysis first detects and removes values which are impossible – for example, zero readings from a sensor (which may arise, for example, because a nurse removed it in order to provide care).

Further artifact removal is carried out by autoregressive (AR) modelling. This flags all values outside a dynamically updated acceptance interval and repairs some transient artifacts. AR coefficients were learnt using the biosig toolbox,[3] using a separate NICU dataset where artifacts had been marked up (courtesy of J. Quinn [39]).

[3] http://biosig.sourceforge.net/.

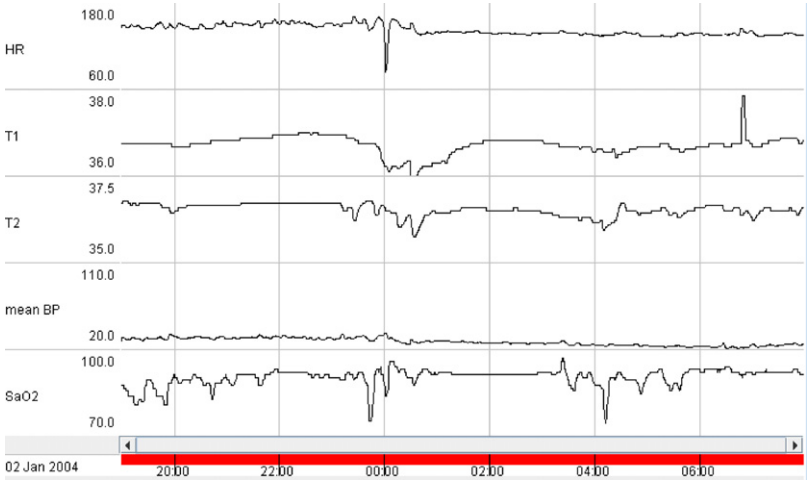

Fig. 5. Example physiological data sampled once per second.

The module now looks for specific patterns of interest in the NICU context, such as bradycardias and desaturations. After a comparison of different techniques [40] we implemented a thresholding method to detect such events, together with an estimate of the baseline using a median filter (window size = 15 s) to find the start and end times. BT-Nurse also computes the number of bradycardias and of desaturations, categorising these as *none*, *few*, *some* or *many*. It also looks at whether desaturations and bradycardias occur at the same time.

Having worked with short time scale events, the module now establishes the longer term characteristics of the signal. To avoid getting bogged down in too much detail, a second median filter is applied with a much larger window size (5 min).

First a bottom-up segmentation [41] is carried out to create a piecewise linear approximation to the signal. The algorithm works by first constructing a very detailed linear approximation to the data, and then repeatedly merging similar adjacent linear segments until there are no adjacent segments which are similar enough to be merged. To adapt the segmentation to the dynamic of the signals, the tolerance error thresholds are based on the variance of the (median-filtered) data; this variance is multiplied by constants which are determined empirically. The individual segments become event instances in the ontology.

Finally, the data for the entire shift is analysed and categorised for each sensor in a number of ways. If at least 75% of the values fall in one of the pre-specified ranges for that sensor (*normal*, *high*, *low*, etc.) then the values for that sensor for the shift are labelled with that range. The *highest* and *lowest* values for the sensor are defined as the 95% and 5% centiles of the distribution of values over the shift. The signal is judged to be *stable* if (*highest* − *lowest*) is less than 20% of the range defined by *high* + *normal* + *low*.

3.2.4. Importance of events

BT-Nurse needs to determine the importance of events; this information is used in later stages to determine whether an event should be mentioned in the generated text. For events which are extracted from the signals, the importance is computed in a number of ways.

For bradycardia and desaturation events, the importance is related to duration and depth. The importance of the event is also weighted differentially according to the values reached (i.e. if a value is within a range that warrants serious concern, the event has a higher weight). The greater the number of bradycardias and/or desaturations, the higher is the importance.

For a flat (zero gradient) segment, the importance is related to the range into which its value falls – *very_high* and *very_low* being the most important. For a segment with a positive or negative gradient, the importance is an empirical function of the magnitude of its slope and of its duration. Thus a very steep (but short) segment is important, as is a gradual rise which occurs over a long period of time.

The description of a signal over the entire shift is always judged to be important.

The derivation of the importance of other events will be described in Section 3.2.6.

3.2.5. Data pre-processing

Before any knowledge-based inferencing or summarisation can start, we need to have as accurate and consistent a picture of what has happened to the baby during the shift as possible. We decided that we would represent this as a set of *states* which exist over a period of time (possibly with uncertainty as to the end-points). So, for example, we would want to know that the baby was intubated from time t_1 to time t_2 (this refers to the baby's being in the *state of being intubated*, rather than to the event of actually performing the intubation that brings about this state). This would then allow us to determine, for example, if two states overlapped.

However this is not the way in which the data is presented to us. Often the style of data entry is configured to fit in with current clinical practice. One example of this is that if the baby is intubated at delivery, this fact may be entered into the database twice; BT-Nurse has to recognise this occurrence and delete one of the entries. Another example is the creation of long-term states from point records. For example, *Badger* (in its hourly observations table) records the drugs a baby is receiving at particular points in time, but it usually does not record exactly when drug administration starts or stops. BT-Nurse merges these observations into a single state. For instance, if morphine is being administered at 10:00, 11:00, 12:00 but not 09:00 and 13:00, then BT-Nurse creates a DRUG-ADMINISTRATION state for morphine whose start time is between 09:00 and 10:00, and whose end time is between 12:00 and 13:00.

As we have said, it is inevitable with patient data entry in the real world, that there are errors of various types. Some are simple to correct, for example if the user has used the wrong units and entered a value which is 100 or 1000 times too high. Some are impossible to detect, for example the entry of a numeric value which is incorrect, but nevertheless plausible. There are some cases which we can try to correct. For example, the nurse may have forgotten to enter an extubation. However the ventilator mode is recorded hourly and

Please cite this article in press as: Hunter J, et al. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.09.002

we may observe a transition from CMV, a mode in which the baby *is* ventilated, to CPAP, a mode in which he is *not*. It can therefore be inferred that the 'intubated' state terminated somewhere between these two data points (since intubation is always associated with ventilation).

As well as representing states it is often important for clinicians to be made aware of the changes between states. It is computationally convenient at this point to generate states corresponding to these changes (i.e. states of change). This means that subsequent modules can test for change directly without having to discover changes every time they are required.

3.2.6. Data interpretation

Data interpretation uses medical knowledge to enhance the information recorded in the ontology, in several ways. This knowledge is encoded in a set of *if . . . then . . .* rules executed by a forward chaining inference engine which makes a single pass through the data. We initially used an existing forward chaining rule system [42], but this was discarded in favour of our own rule syntax and interpreter which provided (i) closer integration with the instances in the ontology and (ii) a more explicit representation of temporal relationships. These rules provide four main functions:

Abstraction. This groups a number of (normally) simultaneous events into a single, higher level event. At present this is only used to infer respiratory state (alkalosis, acidosis, etc.) from simultaneous readings of pH, base excess, and CO_2 (all obtained from the on-ward blood gas analyser). In our running example the *respiratory_acidosis* rule:

IF	pH is *low* or *very_low*
AND	base_excess is *normal*
AND	CO_2 is *high* or *very_high*
THEN	the baby has *respiratory_acidosis*

fired at 23:09 and 06:15.

Importance. Recall that importance estimates the medical significance of an event, and is used when determining whether the event should be mentioned in the text. For example, an endotracheal (ET) suction is a fairly routine event. However if the suction yields blood-stained secretions then its importance is increased considerably.

Relations between events. The types of relation of interest are: *association*, *contains*, *precedes*, *causes*, *enables*, *triggers* – see Section 3.2.1 for definitions. Relations between events are very important for generating high-quality texts. Very often a rule does not require the events to be simultaneous, but imposes some temporal constraint between them (usually proximity and/or order) if the link is to be established. Examples of the kind of knowledge expressed in these rules are:

- If any drug is administered close to an intubation, the process of intubation is considered to 'contain' the administration.
- The administration of caffeine before an extubation is associated with that extubation.
- The administration of surfactant (a drug) can cause bradycardia and/or desaturation.
- The administration of surfactant is (hopefully) associated with a decrease of FiO_2 (percentage of inspired oxygen) in the long term.
- An intubation can cause step changes to any channel associated with ventilation.
- A significant increase in FiO_2 may trigger an intubation.

In our running example, the administration of morphine and suxamethonium at 00:15 is linked to the intubation at the same time.

Potential problems. These rules are the nearest that BT-Nurse comes to giving advice, in that they point out where the current state of the baby might give rise to future concerns/actions. For example:

- If the baby is intubated and an ET suction has produced large amounts of sectretions, then there is the possibility of the ET tube becoming blocked.
- If the last intubation was not followed by a blood gas sample then one should be taken.

In our running example, the ET suction at 05:28 yielded purulent secretions which gives rise to an increased risk of infection.

3.2.7. Document planning

Document planning determines the content and structure of the generated text using a mixture of techniques. For some sections of the text, such as Current Status (see Fig. 1), schemas [43] are used; these are essentially fixed structures populated by relevant events. For example, the first paragraph of Current Status always gives the ventilation state of the baby at the time of handover; the document planner pulls this information out of the ontology and inserts it into the document plan.

For other sections, especially *Events During the Shift*, document planning is more dynamic. The key challenges are deciding which events to include in the text (which can only mention 100 events at most, out of the thousands produced by the above processes), and how the reports of events should be ordered. We decided to base selection on a key event algorithm [44]. Basically the document planner identifies a small number of key events during the shift and generates a paragraph around each of these key events; these paragraphs typically describe related events (as determined by the data interpretation module) as well as the main event itself. This empirically worked better than the alternative of simply selecting the 100 most important events. Ordering can in principle be done in many ways, including by body system, by time, by importance, and by relationships. It was decided to first order sections by body system (e.g. respiratory), and then order key events by time (not importance). Within a paragraph, ordering is based on relationships between the key event and other messages. Thus, within a specific section related to a body system, paragraphs are headed by key events which are consecutive; temporal shifting can however occur within paragraphs.

For example, Fig. 6 shows the document plan for the underlined sentences in the following paragraph from the *Events During the Shift* section of our example:

> The baby was intubated at 00:15 and was on CMV. Vent RR was 50 breaths per minute. Pressures were 20/4 cms H_2O. FiO_2 was 29%. Tidal volume was 1.5. He was given morphine and suxamethonium. MAP was raised from 6 cms H_2O to 8 cms H_2O.

In this case, the document planner identifies the intubation event as a key event, based on the importance value assigned during data interpretation. It also finds a number of related events identified by the data interpretation module; in particular, the intubation caused a shift to a new ventilation mode (CMV), which included new settings to several ventilator parameters, including respiratory rate and several pressure settings. As part of the intubation procedure, the baby was given two drugs, morphine and suxamethonium.

The document planner also performs various narrative optimisations, to address narrative deficiencies which had been identified in the evaluation of the BT-45 system [7]. In particular, it deals with the *continuity* problems which arise when only some trends in a channel are described; this confuses readers. An example from BT-45 (italics added) is as follows

> There were 3 failed attempts to insert a peripheral venous line at 13:53. *$TcPO_2$ suddenly decreased to 8.1*. SaO_2 suddenly increased to 92. *$TcPO_2$ suddenly decreased to 9.3*.

Users complained that it made no sense for $TcPO_2$ to decrease to 9.3 when the last value mentioned for this parameter was 8.1 (and

Please cite this article in press as: Hunter J, et al. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.09.002

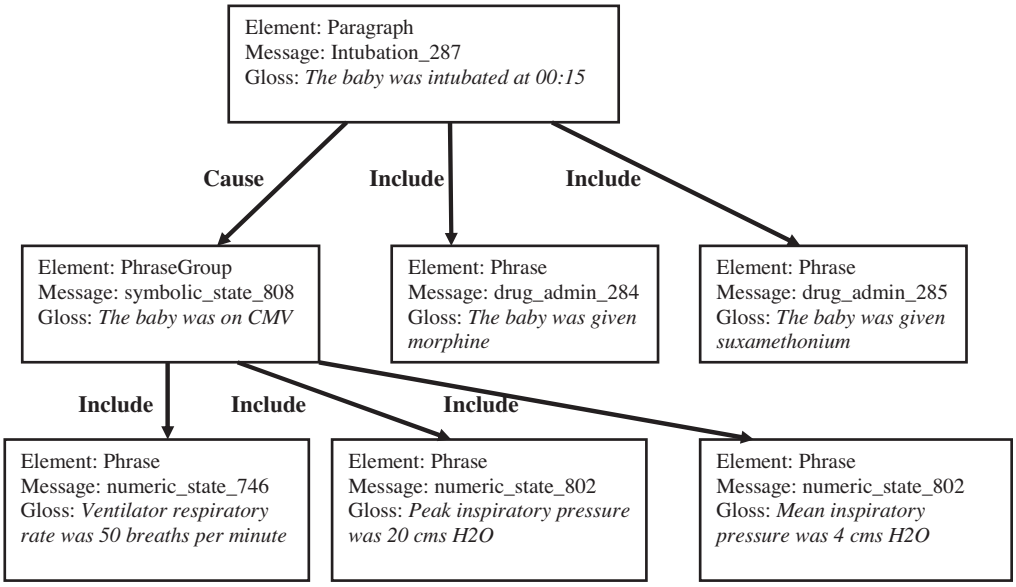

Fig. 6. Part of a document plan.

decreasing). The problem arose because the document planner did not describe the intermediate trend (where $TcPO_2$ went up to 19), because this had a lower importance value.

The BT-Nurse document planner addresses this issue by keeping track (in its discourse model) of the most recent information communicated about each data channel. When it detects a possible continuity problem, it adds an extra *after* phrase to describe the behaviour of the channel in the intermediate period. Using this strategy, the above information would be expressed as follows:

> There were 3 failed attempts to insert a peripheral venous line at 13:53. *$TcPO_2$ suddenly decreased to 8.1*. SaO_2 suddenly increased to 92. *After increasing to 19, $TcPO_2$ suddenly decreased to 9.3*.

As discussed in the evaluation section, this strategy seemed successful; there were many complaints about continuity issues in the BT-45 evaluation, but none in the BT-Nurse evaluation.

3.2.8. Microplanning and realisation

The outcome of document planning is a tree whose nodes are ontology instances representing events or states and whose edges are labelled with the type of relation between the events. The Microplanner traverses this document plan and performs three main tasks, namely *Content Selection*, *Aggregation* and *Temporal Planning*. As a result of these stages, the document plan tree is converted to a set of semantic representations; these are then passed on to a realisation module which maps them to linguistic structures and performs tasks such as constituent ordering and inflectional morphology. Realisation is carried out using the SimpleNLG engine [45]. Since it is a relatively straightforward process, we shall not discuss it further in this section.

Content Selection uses rules to map each ontology instance to a semantic structure depending on its ontological type. In the resulting representation, each event or state is mapped to a linguistic predicate and its relevant properties are mapped to arguments of that predicate. The required arguments are specified in the predicate's lexical entry. For example, the INTUBATION event at the root node of the document plan subtree discussed in the previous section has the predicate (verb) *intubate*, and this requires an *agent* (the person/s performing the action) and a *patient* (in the generic, linguistic sense – i.e. the undergoer of an action). The rules in this case specify that the latter should point to the ontology instance representing the baby about whom the summary is being generated, while the *agent* is null, since it is always assumed that the action is being performed by professionals whose precise identity is usually not specified in the data (and is in any case irrelevant). Predicate arguments which are themselves ontology instances (e.g. the *baby* instance which is the patient of *intubate*) are subject to a second content selection stage, in which the microplanner selects the content for noun phrases. Based on what has been expressed so far, it is possible that an entity be referred to using a pronoun, if it was mentioned previously in a sufficiently salient position in the discourse to permit the reader to resolve the pronoun [46]. For example, after the initial mention of the baby in the third paragraph of the *Events during the shift* section of Fig. 1, a later mention in the same paragraph is simply rendered as *he*.

The second task carried out by the microplanner is *Aggregation*. While many NLG systems perform aggregation of syntactic structures [47], the BT-Nurse microplanner performs it at the semantic level, using the rhetorical relations that label the document plan tree edges to determine whether two events should be merged into one. This is once again performed through a small set of rules which fire in response to (a) a specific rhetorical relation (e.g. the CAUSE relation holding between the intubation and the CMV state in the document plan tree example); (b) the types of events related. If a rule's preconditions are satisfied, the two events are merged and will eventually be realised as a single sentence, consisting of two clauses which are coordinated or subordinated, depending on the relation. In the present example, coordination using *and* yields the sentence: *The baby was intubated and was on CMV*. Here, simple coordination is sufficient to express the causal relation, since it is always clear to a nurse that CMV ventilation is used immediately following intubation. In other instances, more explicit connectives (such as *because* are used).

Following aggregation, *Temporal Planning* determines how an event is to be linguistically located in time. While paragraphs in a document are temporally ordered by key event, there is no guarantee that the key event which is mentioned first in a paragraph will be the first one which actually occurred. Thus, allowing a reader to identify the relative time of events has to rely on linguistic mechanisms, especially tense and temporal modifiers.

Tense is a well-studied topic in formal and computational semantics [48–50]. Many accounts follow Reichenbach [51] in viewing it as an anaphoric category, whereby an event has both an *event time* (ET) and a *reference time* (RT). ET is always determined

by the microplanner based on the actual time interval specified for an event in its corresponding ontology instance. ET is primarily responsible for determining whether an event or state should be described using the present or the past tense, by comparing it to the present time (the system's clock time when it generates the summary; this is sometimes referred to as the *speech time* (ST), in semantic accounts). Present tense is only used in the *Current Status* sections, where a baby's state at present is described. Here, the present is used for states which remain true when the summary is generated (i.e. for those states where ET = ST); for all other events, the tense is always past (since ET < ST).

Even if an event is known to be in the past, the precise tense to be used (e.g. a simple tense such as *was intubated* or a perfect tense such as *had deteriorated*) depends on the relationship between ET and RT. RT is determined based on two factors: the *discourse mode* and the *temporal focus*. The microplanner distinguishes between two discourse modes [52]. In the *Current status* section of a summary, the discourse mode is *deictic*, that is, all events and states are related to the present time, giving rise to a relatively static description of states of affairs. Thus, for all events and states in this section, RT = ST. By contrast, the *Events During the Shift* section is intended to give the reader a more dynamic picture, where events succeed each other in time. Here, the RT of an event within a paragraph is inherited from the previously mentioned event; thus, for any two events i and j, where j is mentioned immediately after i, $RT_j = ET_i$. This also helps to determine the tense: if, for event i, RT > ET (so that this event actually occurs before the previously mentioned event), a perfect tense is used. An example of this is shown in the second paragraph of the *Events During the Shift* section of Fig. 1, where the second event is described using a perfect tense (*had deteriorated*) because it had occurred by the time the previously mentioned event (the ABG) was complete.

The last example also serves to illustrate the notion of *temporal focus*. Following Moens and Steedman [53], we model events (represented as intervals) in BT-Nurse as consisting of an initial phase, a nucleus and a final, consequent state. For example, an interval during which blood gas parameters deteriorate starts from an initial phase where they have not yet deteriorated, and ends in the consequent state where they have. The temporal focus in the microplanner, which is determined by the type of event or state, specifies which of the three main phases of an interval are relevant. In the example under discussion, the focus is on the consequent state (the point where the parameters have deteriorated). Thus, the ET of this event is determined to be its *end time* and the relation between ET and RT is determined accordingly.

Apart from tense, temporal planning also involves the selection of temporal modifiers. Here, the strategy is relatively simple: all key events are expressed using an absolute time phrase (such as *at around 19:45*), where times are rounded off to the nearest quarter hour if necessary and expressed using *around* to convey this. Temporal modifiers are sometimes used with consequent states, as in the case of the blood gas deterioration, where the fact that it is the end point of a longer interval that is in focus is conveyed by the use of modifiers such as *by about 06:15*.

3.3. System development

3.3.1. Knowledge acquisition

There were two types of knowledge which had to be acquired:

(i) The information to be contained in the summary and its presentation
(ii) The underlying medical knowledge required for data interpretation.

We attempted to acquire both types of knowledge through corpus analysis. Clinical staff involved in the project wrote 32 example summaries of a shift using historical data; in what follows, we refer to this as our *corpus*. Its limited size was primarily due to the considerable amount of time and expertise required to sift through a large (several megabytes) dataset related to a single shift, to identify the important elements, and to write a report in English. Furthermore, the corpus itself was intended as a set of exemplars to be analysed manually (rather than, say, subjected to machine-learning or data mining techniques).

Fig. 2 is an extract from one of these summaries. The corpus authors were asked not to make diagnoses or other complex inferences, but instead to focus on selecting and summarising key events, and presenting them in a well-structured English text.

Constructing an artificial corpus was necessary because nurses do not habitually write shift summaries of this size. Indeed, our practical purpose in designing BT-Nurse was to address the potential knowledge transfer gaps that current practice might cause. We analysed the corpus texts to determine what information was communicated, how texts were structured, and what language was used. One challenge in performing this analysis was that the corpus texts were very diverse in terms of length, structure, and information presented; some diversity is expected in this kind of exercise, but the diversity in the BT-Nurse corpus was unusually large. We discussed this with the corpus authors, who explained that there is no standard for the format or detail that should be contained in a natural language shift handover document and that the diversity in the corpus texts reflected the diverse range of patients and illness severity represented.

The corpus texts were also used as the starting point for acquiring the underlying medical knowledge (e.g. the definition of *respiratory acidosis*). Such an approach is in contrast to structured sessions with experts of the sort described by Scott et al. [54]. It is arguable that a more systematic approach would have allowed us to explore more unusual cases which did not appear in our limited corpus.

3.3.2. Data acquisition

Although BT-Nurse was ultimately tested on unseen live patient data (see Section 4) there was a clear need for historical data for development and test purposes. Given that data and developer were on opposite sides of the hospital security firewall, permission was sought and granted to export a sample of patient data provided that it was fully anonymised. All proper names and numbers that could be used to identify a person (i.e. baby, parent, relative, friend, nurse, doctor, etc.), organisation (e.g. hospital), identifying number (e.g. telephone number, hospital ID) or place (e.g. room in hospital, town, area) were replaced by null or anonymous identifiers. Fields which contained only a name or number were easy to anonymise and the field was set to null.

Fields containing free text were more difficult. A lookup table (which remained at the hospital) provided a mapping from real names to anonymised identifiers, which simply allowed us to keep track of the role of a person (e.g. doctor, nurse) and their seniority. The anonymisation software scanned free text for full or partial matches to names in the lookup table; the partial matches (within a pre-set edit distance threshold) accounted for some of the misspellings of names. Patterns were also used to match addresses, telephone numbers and other potentially identifying information; these too were replaced with placeholders (such as TEL or ADDRESS). In addition to these anonymisation methods, we also applied a statistical named entity recogniser,[4] which had been

[4] From the open-source LingPipe library: http://alias-i.com/lingpipe/.

trained on legacy NICU data from another project. This was used to detect potential named entities which had been missed by our heuristics. Given the limited amount of training data available, the likelihood of false positives in the statistical named entity recognition was quite high; thus, all output from the recogniser was checked manually by one of the researchers. Once (semi-)automatic anonymisation was complete, a sample of anonymised free text was extracted and checked with a senior neonatal nurse, for possible errors. This process does not absolutely guarantee that all identifiers were replaced in free text but it was the best that could be done given the volume of data.

In addition to anonymising free text and named fields, timestamps in the database were shifted to a common reference point, so that all babies in the development database had the common birth date of 1st January 2004. This shifting improved anonymity, but it still allowed us to identify the correct time of day of an event recorded in the database, as well as the period that elapsed between the patient's birth and the event itself. Apart from timestamp fields within the database, dates in free text comments were simply replaced by a DATE label. Since the time series data recorded by *Badger* is stored separately from the main database, dates in the physiological data files (including the names of the files) were also shifted as described above.

3.3.3. Data interfacing

BT-45 (our earlier system) and BT-Nurse differed in a number of respects, especially the period covered – 45 min vs. 12 h – and purpose of summary – decision-support vs. shift handover. However, by far the biggest difference as far as development was concerned was that BT-45 used discrete data collected by a research nurse under controlled experimental conditions, whereas BT-Nurse used only that data which was routinely entered and available in the electronic record and operated under the (self-imposed) constraint of not being able to ask any further questions of the user.

The *Badger* system was clearly designed primarily to facilitate data entry and display for human users; the way in which data is entered is often configured to fit in with current clinical practice. *Badger* was not designed to generate what BT-Nurse needed, i.e. a set of states, with accurate timings, over which specific propositions held (e.g. the baby is intubated, a specific drug is being administered).

Thus, establishing the start and end times of states (given that *Badger* does not ask for event times) was not trivial and we were often reduced to setting fairly wide ranges on these times. Absolute times on their own may not be too important, but if one is trying to establish the relative times of two events (did event *A* occur before event *B* or vice versa) and if absolute times are the only times available, then accuracy becomes significant.

The problem is compounded by the fact that human data entry is subject to the common errors of omission (forgetting to enter the data) and incorrectness (entering the wrong value). Many of these errors are unimportant for human users. The nurse who is closely involved in caring for the baby can usually spot inconsistencies very quickly. For example, if it has not been recorded that the baby has been extubated, a quick glance will confirm that extubation has taken place. However these errors pose considerable problems for a software system such as BT-Nurse which was wholly dependent on the recorded data.

Some grounds for optimism here may come from the increased interfacing of external equipment. For example, on our NICU, the ventilator had been interfaced to *Badger* on an experimental basis; however the results were not routinely available.

A further difficulty comes from the fact that some of the information is entered as free text; this is perfectly acceptable for subsequent re-display but complicates the task of information extraction.

Some of the functionality of the data pre-processing module is generic – for example horizontal interpolation (as defined by Shahar and Musen [55]), discussed above in relation to drug administration. However most of the functions implemented in the data translation and data pre-processing modules had to be designed to meet local circumstances which will almost certainly vary from unit to unit.

3.3.4. System deployment

Linking BT-Nurse to the existing *Badger* system was surprisingly easy thanks to excellent relations with Clevermed (the system developer). A new tab on the cot-side *Badger* system was provided to generate a request to BT-Nurse for a summary (in the form of a web-page). BT-Nurse, running on its own experimental server, interrogated the live database, generated the summary, and delivered it back to the cot-side for display. Summaries were generated in less than 1 min, with no noticeable impact on other users of the *Badger* system.

4. Results

In general terms, there are three approaches to evaluating data-to-text systems such as BT-Nurse [56]:

- *Compare computer-generated texts to human-written corpus texts based on the same input data*. This is often done using automatic metrics which compute average *n*-gram overlap, such as BLEU [57] or ROUGE [58], or measures which focus on content elements, such as the Pyramid method in summarisation [59]. Such methods usually require several examples of 'gold standard' reference texts based on the same input data. An alternative is to perform a detailed analysis, for example by a discourse analyst; we have reported the results of such a study elsewhere [60].
- *Evaluate the impact of a text on users*. Again there are many variations, ranging from controlled psychological experiments in artificial contexts [5] to randomised controlled clinical trials [26].
- *Ask human experts to rate the quality of texts*. Again there are a variety of techniques, including asking subjects for Likert-scale ratings of different aspects of individual texts [61], and asking subjects to compare alternative texts and state which they prefer [16].

There are several reasons why the first type of evaluation (using automatic metrics) was not appropriate in the present context. First, such comparisons usually rely on the existence of a gold standard corpus to which system-generated texts can be compared. In the present case, such a gold standard was not available. Although a number of expert texts were used in the course of the development of BT-Nurse, the resulting corpus was of limited size, with only one reference text per input dataset; small reference corpora tend to compromise the reliability of automatic metrics. As noted above, a larger corpus was not feasible, given the enormous amount of time and effort that it takes an expert to write such summaries. A more serious problem was that the task that BT-Nurse was designed to perform was not one that is routinely carried out by nurses on the ward; this puts the very existence of a gold standard into question. Indeed, the rationale behind the collection of our initial corpus was to obtain an example set of texts based on historical data that could serve as a starting point for system design in the absence of actual examples; however, we had no a priori guarantee that texts in this corpus would be considered 'optimal' by a majority of users. Finally, recent work on comparative evaluation in NLG has suggested that evaluation results obtained from automatic metrics may have little or no relationship with outcomes obtained using methods involving human judges or studies measuring impact on

Please cite this article in press as: Hunter J, et al. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.09.002

Table 1
Nurses' views of the BT-Nurse summaries (% of trials).

	Understandability			Accuracy			Helpfulness		
	Agree	Neutral	Disagree	Agree	Neutral	Disagree	Agree	Neutral	Disagreed
Incoming	92.4	7.6	0	73.9	23.9	2.2	56.5	35.9	7.6
Outgoing	87.7	8.2	4.1	65.8	24.7	9.6	61.6	30.1	8.2
Overall	90.3	7.9	1.8	70.3	24.2	5.5	58.8	33.3	7.9

users [56,62]; similar concerns have been raised in other areas of NLP [63,64].

An evaluation of the second type, such as a clinical trial measuring the impact of texts on users, would be highly desirable, but was not considered to be feasible in the present context, because substantial progress would be needed on the issues identified below (such as dealing with imperfect input data).

Based on the above considerations, we opted for a study of the third type, in which human experts (nurses) were asked to rate generated texts related to patients in their care. We used a version of BT-Nurse which was deployed on-ward and generated partial shift summaries (background information, problems, respiratory status, and cardiovascular status) from data collected during the shift. Time constraints within the project meant that it was not possible to develop the software to the point where it could generate complete shift summaries (which would, as a minimum, have also included sections on thermal control and fluids/feeds).

The main BT-Nurse evaluation was done by asking nurses to rate BT-Nurse texts for understandability, accuracy, and helpfulness. Since this was done on-ward, nurses saw BT-Nurse texts for babies they were currently looking after. This meant that the test (evaluation) data was also different from the training data.

4.1. Methodology

About 1 h before the end of a selected shift, our researcher (herself a highly experienced neonatal nurse) requested a summary. For ethical reasons, the researcher had the ability to delete any inaccurate content in the BT-Nurse text which might potentially confuse the nurse and adversely impact her care of the baby. However in practice during the evaluation, the researcher made no such deletions.

After reading the summary, the outgoing nurse in charge of the baby was asked to indicate agreement, disagreement or neutrality with respect to the following three questions:

- "The BT-Nurse summary was easy to understand" [understandability]
- "The BT-Nurse summary is accurate" [accuracy]
- "The BT-Nurse summary would help me write a shift summary" [helpfulness]

The nurse and the researcher were then invited separately to enter free text comments on any aspect of the summary.

Summaries for the incoming nurses were generated after the end of the shift and the incoming nurse was given enough time for the normal handover before being shown the summary. The incoming nurses were asked the same questions as the outgoing, except that the final question was:

- "The BT-Nurse summary would help me in care planning" [helpfulness]

In addition the incoming nurse was asked to record if she had:

- had an oral handover;
- looked at the *Badger* charts;
- looked at the *Badger* Nurse Shift Summary.

Again, the nurse and the researcher were invited separately to enter free text comments on any aspect of the summary. Note that the incoming nurse had access to sufficient information to enable her to form some opinion as to the accuracy and helpfulness of the summaries.

In short, BT-Nurse was evaluated by nurses who read summaries about real babies under their care, while they were on the hospital ward, and in the context of a real task (shift handover) which they needed to do.

The evaluation was conducted over a 2-month period, after an initial 1-month pilot phase. If we define a 'trial' as the evaluation by one nurse of the summary for one shift, we conducted 165 trials – 73 with outgoing and 92 with incoming nurses. A total of 148 summaries were generated for 31 individual babies. In 131 cases a summary was seen by only one nurse (outgoing or incoming); in the remaining 17 the summary was seen by both nurses. On average, each baby was seen by 2.3 nurses (maximum 6). Of a nursing staff complement of 93, 54 different nurses participated. On average each nurse saw 4 different babies; only 4 nurses saw more than 8 different babies.

4.2. Information sources consulted by outgoing nurses

Of the 92 incoming nurse trials, only one nurse reported not receiving an oral handover from the outgoing nurse. In 84 of the trials (91%) the nurse reported looking at the on-line *Badger* charts, but in only 54 (59%) did she report reading the *Badger* Nurse Shift Summary. We could interpret this last figure (which is quite low) in one of two ways. Perhaps the nurses do not like reading summaries of any kind – in which case it might be difficult to get them to read automatically generated ones. Alternatively, they might find the existing *Badger* summaries somewhat uninformative – in which case a better presented summary might be consulted more often.

4.3. Responses to questions

The detailed results for the questions relating to understandability, accuracy and helpfulness are given in Table 1 where the averages are over trials. We also averaged the responses over individual nurses before performing statistical tests. The results are similar.

As Table 1 shows, the majority response was positive for all three questions that the nurses were asked. We compared response frequencies for each of the categories in each of the three questions, using a χ^2 goodness of fit test. Overall, there were significant differences between the number of positive, negative and neutral responses for all three statements: *understandability* ($\chi^2 = 241.89$; $p < 0.001$); *accuracy* ($\chi^2 = 110.22$; $p < 0.001$); and *helpfulness* ($\chi^2 = 64.15$; $p < 0.001$). This suggests that the majority of positive responses we observe on all three questions is statistically reliable.

The response frequencies in Table 1 suggested that there may be differences between incoming and outgoing nurses in the likelihood of certain responses. For example, with the exception of the *helpfulness* question, there seems to be a greater tendency to

Please cite this article in press as: Hunter J, et al. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.09.002

Table 2
Percentage of positive responses to questions as a function of summary length.

Length	*N*	Understandability	Accuracy	Helpfulness
60–110	44	93.2% (41)	77.3% (34)	63.6% (28)
111–160	54	88.9% (48)	68.5% (37)	61.1% (33)
161–210	32	93.8% (30)	78.1% (25)	56.3% (18)
211–260	20	85% (17)	55% (11)	55% (11)
261–310	12	100% (12)	66.7% (8)	50% (6)
>310	3	33% (1)	33% (1)	33% (1)

reply positively among incoming nurses. These observations raise the question of whether nurse role (incoming/outgoing) may have had a significant impact on responses. We tested this using a linear mixed effects analysis, for each of the three questions asked. In each case, the model included role as fixed effect, with random intercepts and slopes for nurses and patients. There was no main effect of nurse role on response type in any of the questions: *understandability* ($Z = 0.03$, $p > 0.5$); *accuracy* ($Z = 1.04$, $p > 0.3$); or *helpfulness* ($Z = -1.68$, $p > 0.5$). In short, the apparent differences between incoming and outgoing nurses are not statistically reliable.

Another factor which may impact the proportion of subjective positive responses for any of the three questions is the severity of a patient's illness. This is reflected in the level of care of the patient in question, classified according to the British Association for Perinatal Medicine (BAPM) classification. The BAPM scoring system was devised to quantify resources required by UK NICUs in relation to the severity of a patient's condition. It distinguishes four levels: 1 (intensive care), 2 (high dependency), 3 (special care) and normal care. In our sample, there were 114 trials with the patient in intensive care, 37 in high dependency, and 14 in special care. To test whether there was an impact of severity of illness (as reflected by BAPM), we incremented the above models, factoring in BAPM as a fixed effect in addition to nurse role. This allows us to test both whether BAPM itself impacts responses, but also whether it interacts with role (for example, because incoming or outgoing nurses alter their views as a function of perceived severity of illness). As before, there was no main effect of role; there was also no main effect of BAPM and no interaction, on either of the questions (all Z-values < 1; all p-values > .5). A log likelihood test showed that adding BAPM to the initial model with only nurse role as fixed effect did not improve goodness of fit to the data for any of the three questions (*understandability*: $\chi^2 = 0$; *helpfulness*: $\chi^2 = 10.1$, $p > 0.8$; *accuracy*: $\chi^2 = 14.29$, $p > 0.5$).

The final factor that might have impacted nurses' responses was the length of summaries, which differed among scenarios and partially depends on how many noteworthy (as judged by BT-Nurse) events occurred in a shift. We used Microsoft Word to obtain the number of words in the most important part of the summary (the Respiratory Support and Cardiovascular System sections). This part of the summaries ranged in length from 60 to 436 tokens, with 98.2% falling within the 60–310 token range. Table 2 shows the frequency of positive, negative and neutral responses as a function of word length within this range. To facilitate display, the table only summarises data within the 60–310 range; this is divided into bins at increments of 50 tokens, with the number of summaries and the proportion of positive ("agree") responses to each of the three questions in each bin. To analyse the impact of length on response, we again performed a linear mixed effects analysis, with word count as fixed effect and random intercepts and slopes for participants and items, as before. Note that word count was included as a continuous variable in this analysis. The model on the entire data set showed no main effect of word count on *understandability* ($Z = -0.09$, $p > 0.3$), *accuracy* ($Z = -0.06$, $p > 0.9$) or *helpfulness* ($Z = 0.56$, $p > 0.5$). Excluding texts whose length was below 60 tokens or greater than 310 tokens yields roughly the same results.

Table 3
Dimensions and categories used in the annotation of segments, with mean pairwise agreement and frequencies.

Dimension	Mean pairwise agreement	Categories	Frequency
Overall	0.83	Positive	89% (31)
		Negative	11% (4)
Content	0.73	Missing-content	59% (109)
		Wrong	25% (46)
		Unnecessary	11% (20)
		General	4% (7)
		Good	3 (2%)
		Misplaced	0% (0)
Language	0.66	Poor	100% (11)
		Good	0% (0)

4.4. Analysis of free-text comments

Free text comments were manually segmented, so that each segment addressed one specific aspect of the summary. This gave rise to 237 segments (125 for outgoing nurses and 112 for incoming); we did not include the comments from the research nurse in this analysis. These segments were annotated independently by three of the authors (JH, ER and AG) along several dimensions; the ones which are relevant for the present analysis are summarised in Table 3, together with the categories for each dimension. Table 3 displays the mean pair-wise agreement (using Cohen's κ statistic) between annotators on each dimension. Using standard thresholds for kappa [65], these values suggest good agreement in the *overall* dimension, and tentative agreement in the *content* and *language* dimensions. In the following analysis, we used the majority categorisation for each segment (that is a categorisation agreed upon by two out of three annotators) whenever one exists, omitting the 20 segments for which there was no majority categorisation. Table 3 also shows the frequencies of each category, using the results of the above process.

The results of the analysis can be summarised as follows. First, most segments concerned the *content* of the summary (185). Most of these (109) noted *missing* content. Many of these referred to information (for example about nutrition or weight) which was not presented in the partial shift summaries produced by BT-Nurse. However, even if we disregard such segments, requests for more content were much more common than requests for less (of which there were 20). This suggests that BT-Nurse was under-reporting. As an example, one nurse wrote that the "baby ... is VERY small and ... it should be pointed out that the ETT is size 2.0". BT-Nurse never reports ETT size as usually this is not important; however in some cases (e.g. very small babies) it is important and should be reported.

There were 46 segments concerning *incorrect* content; most of these were due to bugs in the software, but some were due to problems in input data, such as those sensor artifacts which are hard to identify, or incorrect information in the patient record. For example, BT-Nurse sometimes listed current problems which had in fact been rectified; this was because of errors in reading the relevant database tables.

There were only 11 segments about *language*, all of which were negative. Most of these criticisms reflected either bugs or personal preferences of individual nurses. Given the small number of such segments and the overwhelming positive view that the summaries were understandable (90%), we consider that nurses considered 'good' language to be the norm and only commented in exceptional cases where this was not achieved.

Among the 35 *overall* segments, the majority expressed positive views of the summaries; for example: "Very good; easy to understand and is an accurate reflection of today's shift" and "Summary

Please cite this article in press as: Hunter J, et al. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.09.002

is a very useful tool for a quick report about your baby past and present". Four of the overall segments concerned deficiencies in BT-Nurse texts from a high-level *narrative* perspective; for example, not describing causal links between observations and interventions ("The information is correct, but for a shift summary I would want to be able to include reasons for the facts that are presented. For example, 'Temp Gap ranges from 2.1 to 4.2.' I want to add that this was when the baby stopped phototherapy."), or not adequately describing the overall *big picture* ("The above comments are accurate statements, however they do not present a 'picture' of current condition. For example she was re-intubated but we are not told why."). These reflect a more fundamental limitation of current data-to-text technology, which we discuss below.

5. Discussion

5.1. Overview

In the light of the questions posed in Section 1, we consider BT-Nurse to have been generally successful. A very complex system, which accessed voluminous heterogeneous data in a near real-time environment, ran without failure on the ward for several weeks. In the evaluation, an overwhelming majority of summaries was rated by the users as understandable, and a significant, if somewhat smaller, majorities were rated as accurate and helpful. Note that only 5.5% of the summaries were positively rated as *in*accurate and only 7.9% as *un*helpful. We consider that the helpfulness of the summaries implies that the *data selection and abstraction was relevant* and the understandability implies that *the language was fluent and easily readable*.

Moreover, there were some very encouraging comments about how BT-Nurse summaries helped incoming nurses in particular – for example: "Looking at this at handover would help me think about what I had to do for the rest of the day and make a plan" and "BT picked up the change in HR trend that I had not noticed."

5.2. Outstanding issues

Our analysis of the nurses' comments suggests that the biggest problem was missing information. From an algorithmic perspective, this is primarily a problem with determining the importance of events (Sections 3.2.4 and 3.2.6) and is perhaps due to the fact that our importance rules were insufficiently sensitive to context (what is important depends on the baby's medical condition) and also to nurse's work flow (nurses are especially interested in information which affects care planning). One reason for this is that our domain knowledge activities focused more on problem diagnosis and other "conventional" reasoning issues, and less on determining importance. Furthermore, some of the importance rules in BT-Nurse were adaptations of rules developed for our earlier BT-45 system, which were based on knowledge acquisition sessions with doctors; such importance rules do not incorporate knowledge of what nurses want to know in order to do care planning.

Although there were relatively few instances where content was incorrect, and none which the supervising nurse felt were serious enough to warrant deletion, incorrect content does remain a major issue because it could adversely impact care. Where BT-Nurse did report such content, it was partly due to incorrect data in the patient record, and partly to problems in the analysis and interpretation of noisy data. It is possible that our approach to analysis and interpretation was too "aggressive", in the sense that a more conservative approach could have led to fewer incorrect inferences, at the expense of fewer correct inferences as well. A more practical source of content-related errors was the insufficient time spent debugging BT-Nurse on-site at the hospital. Although we spent several months debugging BT-Nurse off-site (at Aberdeen University), on-site debugging (with live data) was limited to the pilot period of 1 month, and many mistakes only showed up on-site (especially mistakes in handling unusual or boundary cases). In the current BT-Family project, we have allocated 6 months (instead of 1 month) for on-site pilot work (including debugging).

It is worth noting that data quality is getting better, as sensors improve and more data is acquired automatically instead of being entered manually. Dealing with noisy or incorrect data will always be a major problem for medical data-to-text systems but there is reason to expect that it will become easier over time.

It is interesting that there were relatively few comments about language. Outright mistakes in language were mostly due to insufficient debugging on expressing information in a few very specific instances (as mentioned above). Putting these aside, while there were clearly some linguistic deficiencies in the generated texts, there are also linguistic deficiencies in human-written shift summaries. Perhaps it is the case that once linguistic quality exceeds a threshold, people regard it as "good enough" and are not particularly concerned with remaining disfluencies. If this hypothesis is true, the threshold will almost certainly depend on genre and expectations; for example leaflets containing information for patients will probably need to meet higher linguistic standards than nurse shift summaries.

5.3. Future work

Looking at BT-Nurse as a whole, we believe that automated summaries are especially useful when the nurses involved in the handover are relatively inexperienced, and hence likely to miss out important information. In part because of this insight, some of us are involved in a new project which is also related to handovers, MIME [66], whose goal is to support handover from first responders to paramedics and ambulance crew. When a medical emergency occurs in a remote area, it may take hours for an ambulance to arrive, and in the meantime the patient may need to be looked after by a community first responder; in other words, by a local volunteer who has a small amount of training and some basic sensors and equipment. These first responders have very limited training (typically just a week or so on a course) and also relatively little experience (a first responder may only deal with a serious case once per year); this amounts to far less training and experience than even the most junior nurse. But knowing what happened to the patient in the first hours after an incident may be crucial to the ambulance paramedics and the hospital doctors who treat the patient later. Hence the goal of MIME is to produce a handover report (based both on data from sensors which the first responder attaches to the patient, and on qualitative information entered by the first responder) which summarises what happened to the patient before the ambulance arrived.

From a more theoretical perspective, we would like to develop better techniques for generating good narratives which include causal links and make the big picture clear. Current data-to-text systems aim to generate reports to present facts; we believe that aiming for a more narrative strategy in generation would enhance the quality and understandability of the text. From a research perspective, this requires more attention to the qualities of narrative that enable it to highlight important information and convey temporal, discourse and causal relations to the reader. Indeed, one could argue that such narrative aspects are perhaps the primary benefits of human-written textual summaries over computer-generated tabular/visual presentations.

We would also like to make better use of free-text comments in the patient record. There is a lot of scope for applying more sophisticated information extraction techniques to pull key information about of the free-text comments. We would also like to

Please cite this article in press as: Hunter J, et al. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.09.002

explore how the data-to-text technology presented in this paper could be integrated with text-to-text summarisation technology, so an integrated summary could be produced of both structured data (using data-to-text) and unstructured free-text comments (using text-to-text).

It is perhaps useful to note that the processing up to and including the data interpretation module is designed to generate a clean description (implemented as instances in the ontology) of the events occurring within the NICU. In our case this description is used to generate a summary, but we believe that the way in which the description is constructed owes little to that specific end use, and could be used as a starting point for other purposes, for example decision-support, adherence to clinical guidelines and clinical audit.

6. Conclusions

BT-Nurse generates partial shift summaries in the Neonatal Intensive Care Unit, focusing on important events related to a patient's respiration and circulation, in addition to background information, current status and potential problems. We believe that our proof-of-concept system and its evaluation suggest positive answers to the research questions we raised at the beginning of this paper: data-to-text technology does appear to be sufficiently robust that a complex system can run on live data for an extended period; it also appears to generate summaries of a quality that is rated as understandable by a large majority of users, and also as accurate and helpful by a significant, if somewhat smaller, majority.

Time constraints prevented us from implementing a complete system, dealing with all relevant aspects of a patient's state, rather than just their respiratory and circulatory systems. Furthermore, our evaluation enabled us to identify some problems related to handling noisy data, knowledge representation and reasoning.

In summary, BT-Nurse has shown that we can confidently foresee a computer system which automatically generates understandable and helpful shift summary texts from a complex, state-of-the-art patient information system holding a large amount of heterogeneous data. Additional development effort would be needed before BT-Nurse could be realistically deployed, or indeed evaluated in a clinical trial which measured patient outcome instead of nurses' perceptions. Nevertheless, we have shown that the technology works, and we believe its long-term potential is considerable.

Acknowledgements

We are grateful to the UK Engineering and Physical Sciences Research Council (EPSRC) for funding the BabyTalk project with grants to the University of Aberdeen (EP/D049520/1) and the University of Edinburgh (EP/D05057X/1). We are also grateful to Peter Badger and Tom Lyon of Clevermed® for always being on hand to answer our queries and to expedite integration of BT-Nurse with the *Badger* system. We thank our reviewers for considered and helpful comments.

References

[1] Cunningham S, Deere S, Symon A, Elton R, McIntosh N. A randomized, controlled trial of computerized physiologic trend monitoring in an intensive care unit. Critical Care Medicine 1998;26:2053–9.
[2] Alberdi E, Becher J-C, Gilhooly K, Hunter J, Logie R, Lyon A, et al. Expertise and the interpretation of computerised physiological data: implications for the design of computerised physiological monitoring in neonatal intensive care. International Journal of Human Computer Studies 2001;55:191–216.
[3] Freer Y, Ferguson L, Ewing G, Hunter J, Logie R, Rudkin S, et al. Mismatched concepts in a neonatal intensive care unit (NICU): further issues for computer decision support? Journal of Clinical Monitoring and Computing 2003;17:441–7.
[4] Law A, Freer Y, Hunter J, Logie R, McIntosh N, Quinn J. 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 2005;19:183–94.
[5] van der Meulen M, Logie R, Freer Y, Sykes C, McIntosh N, Hunter J. 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. Applied Cognitive Psychology 2008;24:77–89.
[6] Hunter J, Freer Y, Gatt A, Reiter E, Sripada S, Sykes C, et al. BT-Nurse: computer generation of natural language shift summaries from complex heterogeneous medical data. Journal of the American Medical Informatics Association 2011;18(5):621–4.
[7] Portet F, Reiter E, Gatt A, Hunter J, Sripada S, Freer Y, et al. Automatic generation of textual summaries from neonatal intensive care data. Artificial Intelligence 2009;173:789–816.
[8] Gatt A, Portet F, Reiter E, Hunter J, Mahamood S, Moncur W, et al. From data to text in the neonatal intensive care unit: using NLG technology for decision support and information management. AI Communications 2009;22:153–86.
[9] Friesen M, White S, Boyers J. Handoffs: implications for nurses. In: Hughes R, editor. Patient safety and quality: an evidence-based handbook for nurses. Rockville, MD: Agency for Healthcare Research and Quality; 2008.
[10] Menke J, Broner C, Campbell D, McKissick M, Edwards-Beckett J. Computerized clinical documentation system in the pediatric intensive care unit. BMC Medical Informatics and Decision Making 2001;1 [Article no. 3].
[11] Strople B, Ottani P. Can technology improve intershift report? What the research reveals. Journal of Professional Nursing 2006;22(3):197–204.
[12] Staggers N, Clark L, Blaz J, Kapsandoy S. Why patient summaries in electronic health records do not provide the cognitive support necessary for nurses' handoffs on medical and surgical units: insights from interviews and observations. Health Informatics Journal 2011;17(3):209–23.
[13] Goddard K, Roudsari A, Wyatt J. Automation bias: a systematic review of frequency, effect mediators and mitigators. Journal of the American Medical Informatics Association 2012;19:121–7.
[14] Reiter E, Dale R. Building natural language generation systems. Cambridge: Cambridge University Press; 2000.
[15] Reiter E. An architecture for data-to-text systems. In: Proceedings of the 11th European workshop on natural language generation. 2007. p. 97–104.
[16] Reiter E, Sripada S, Hunter J, Yu J, Davy I. Choosing words in computer-generated weather forecasts. Artificial Intelligence 2005;167:137–69.
[17] Belz A. Automatic generation of weather forecast texts using comprehensive probabilistic generation space models. Natural Language Engineering 2007;14:431–55.
[18] Goldberg E, Driedger N, Kittredge R. Using natural language processing to produce weather forecasts. IEEE Expert 1994;9:45–53.
[19] Coch J. MULTIMETEO: multilingual generation of weather forecasts. ELRA Newsletter 1998;3(2).
[20] Turner R, Sripada S, Reiter E, Davy I. Selecting the content of textual descriptions of geographically located events in spatio-temporal weather data. In: Applications and innovations in intelligent systems XV. Proceedings of the 27th SGAI international conference on innovative techniques and applications of artificial intelligence. 2007. p. 75–88.
[21] Kukich K. Design of a knowledge-based report generator. In: Proceedings of the 21st annual meeting of the association for computational linguistics (ACL'83). 1983. p. 145–50.
[22] Iordanskaja L, Kim M, Kittredge R, Lavoie B, Polguere A. Generation of extended bilingual statistical reports. In: Proceedings of the 15th international conference on computational linguistics. 1992. p. 1019–23.
[23] Ferres L, Parush A, Roberts S, Lindgaard G. Helping people with visual impairments gain access to graphical information through natural language: the iGraph system. In: Computers helping people with special needs: 10th international conference, ICCHP 2006. 2006. p. 1122–30.
[24] Yu J, Reiter E, Hunter J, Mellish C. Choosing the content of textual summaries of large time-series data sets. Natural Language Engineering 2007;13:25–49.
[25] Hueske-Kraus D. Text generation in clinical medicine: a review. Methods of Information in Medicine 2003;42(1):51–60.
[26] Reiter E, Robertson R, Osman L. Lessons from a failure: generating tailored smoking cessation letters. Artificial Intelligence 2003;144:41–58.
[27] Hueske-Kraus D. Suregen-2: a shell system for the generation of clinical documents. In: Proceedings of the conference of the European chapter of the association for computational linguistics (EACL'03). 2003. p. 215–8.
[28] Afantenos S, Karkaletsis V, Stamatopoulos P. Summarization from medical documents: a survey. Artificial Intelligence in Medicine 2005;33(2):157–77.
[29] Lacaye C, Diez F. A review of explanation methods for heuristic expert systems. Knowledge Engineering Review 2004;19(2):133–46.
[30] McKeown L, Pan S, Shaw J, Jordan D, Allen B. Language generation for multimedia healthcare briefings. In: Fifth ACL conference on applied natural language processing (ANLP'97). 1997. p. 277–82.
[31] Harris M. Building a large-scale commercial NLG system for an EMR. In: Proceedings of the 5th international natural language generation conference (INLG'08). 2008. p. 157–60.
[32] Suominen H, Salakoski T. Supporting communication and decision making in Finnish intensive care with language technology. Journal of Healthcare Engineering 2010;1(4):595–614.

Please cite this article in press as: Hunter J, et al. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.09.002

[33] Mahamood S, Reiter E. Generating affective natural language for parents of neonatal infants. In: Proceedings of the 13th European workshop on natural language generation (ENLG'11). 2011. p. 12–21.
[34] Moncur W, Reiter E, Masthoff J, Carmichael A. Modelling the socially intelligent communication of health information to a patient's personal social network. IEEE Transactions on Information Technology in Biomedicine 2010;14:319–25.
[35] Humphreys B, Lindberg D. The UMLS project: making the conceptual connection between users and the information they need. Bulletin of the Medical Library Association 1993;81:170–7.
[36] Noy N, Crubézy M, Fergusson R, Knublauch H, Tu S, Vendetti J, et al. Protégé-2000: an open-source ontology-development and knowledge-acquisition environment. In: AMIA 2003 annual symposium proceedings. 2003. p. 953.
[37] Allen J. Towards a general theory of action and time. Artificial Intelligence 1984;23:123–54.
[38] Shahar S, Miksch S, Johnson P. The Asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. Artificial Intelligence in Medicine 1998;14:29–51.
[39] Quinn J. Bayesian condition monitoring in neonatal intensive care. Doctoral Dissertation. University of Edinburgh; 2007.
[40] Portet F, Gao F, Hunter J, Sripada S. Evaluation of online bradycardia boundary detectors from neonatal clinical data. In: Proceedings of the 29th IEEE annual international conference of the engineering in medicine and biology society (EMBS'07). 2007. p. 3288–91.
[41] Keogh E, Chu S, Hart D, Pazzani M. An online algorithm for segmenting time series. In: Proceedings of the IEEE international conference on data mining (ICDM'01). 2001. p. 289–96.
[42] Friedman-Hill E. JESS in action: Java rule-based systems. Greenwich, CT: Manning Publications Co.; 2003.
[43] McKeown K. Text generation. Cambridge: Cambridge University Press; 1985.
[44] Hallett C, Power R, Scott D. Summarisation and visualisation of e-health data repositories. In: Proceedings of the UK E-Science all-hands meeting. 2006.
[45] Gatt A, Reiter E. SimpleNLG: a realisation engine for practical applications. In: Proceedings of the 12th European workshop on natural language generation (ENLG'09). 2009. p. 90–4.
[46] Krahmer E, Theune M. Efficient context-sensitive generation of referring expressions. In: van Deemter K, Kibble R, editors. Information sharing: reference and presupposition in language generation and interpretation. Stanford: CSLI; 2002.
[47] Harbusch D, Kempen G. Generating clausal coordinate ellipsis multilingually: a uniform approach based on postediting. In: Proceedings of the 12th European workshop on natural language generation (ENLG'09). 2009. p. 105–44.
[48] Partee B. Some structural analogies between tenses and pronouns in English. Journal of Philosophy 1973;70:601–9.
[49] Webber B. The interpretation of tense in discourse. Computational Linguistics 1988;14(2):61–73.
[50] ter Meulen A. Representing time in natural language. Cambridge, MA: MIT Press; 1995.
[51] Reichenbach H. Elements of symbolic logic. London: Macmillan; 1947.
[52] Caenepeel M. Aspect and text structure. Linguistics 1995;33:213–53.
[53] Moens M, Steedman M. Temporal ontology and temporal reference. Computational Linguistics 1988;14:15–28.
[54] Scott A, Clayton J, Gibson E. A practical guide to knowledge acquisition. Boston, MA: Addison-Wesley; 1991.
[55] Shahar Y, Musen M. Knowledge-based temporal abstraction in clinical domains. Artificial Intelligence in Medicine 1996;8(3):267–98.
[56] Reiter E, Belz A. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. Computational Linguistics 2009;35(4):529–58.
[57] Papineni S, Roukos T, Ward W, Zhu W. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics (ACL'02). 2002. p. 311–8.
[58] Lin C-Y. ROUGE: a package for automatic evaluation of summaries. In: Proceedings of the workshop on text summarization branches out (WAS 2004). 2004.
[59] Nenkova A, Passonneau R, McKeown K. The pyramid method: incorporating human content selection variation in summarization evaluation. ACM Transactions on Speech and Language Processing 2007;4(2).
[60] McKinlay A, McVittie C, Reiter E, Freer Y, Sykes C, Logie R. Design issues for socially intelligent user interfaces: a qualitative analysis of a data-to-text system for summarizing clinical data. Methods of Information in Medicine 2010;49:379–87.
[61] Lester J, Porter B. Developing and empirically evaluating robust explanation generators: the KNIGHT experiments. Computational Linguistics 1997;23(1):65–101.
[62] Gatt A, Belz A. Introducing shared task evaluation to NLG: the TUNA shared task evaluation challenges. In: Krahmer E, Theune M, editors. Empirical methods in natural language generation. Berlin/Heidelberg, Germany: Springer; 2010. p. 264–94.
[63] Calliston-Burch C, Osborne M, Koehn P. Re-evaluating the role of BLEU in machine translation research. In: Proceedings of the 11th conference of the European chapter of the association for computational linguistics (EACL'06). 2006. p. 249–56.
[64] Dorr B, Monz C, President S, Schwartz R, Zajic D. A methodology for extrinsic evaluation of text summarization: does ROUGE correlate? Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarisation. 2005. p. 1–8.
[65] Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Computational Linguistics 2008;34(4):555–96.
[66] Nguyen H, Mellish C, Mort A, Kindness P, Knight J, Reiter E. Using NLG to manage information in medical emergencies. In: Proceedings of digital engagement 2011 – the second digital economy all hands conference. 2011.

Please cite this article in press as: Hunter J, et al. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.09.002