MIME - NLG in Pre-Hospital Care

Anne H. Schneider Alasdair Mort Chris Mellish Ehud Reiter Phil Wilson
University of Aberdeen
{a.schneider, a.mort, c.mellish, e.reiter, p.wilson}@abdn.ac.uk

Pierre-Luc Vaudry
Université de Montréal
vaudrypl@iro.umontreal.ca

Abstract

The cross-disciplinary MIME project aims to develop a mobile medical monitoring system that improves handover transactions in rural pre-hospital scenarios between the first person on scene and ambulance clinicians. NLG is used to produce a textual handover report at any time, summarising data from novel medical sensors, as well as observations and actions recorded by the carer. We describe the MIME project with a focus on the NLG algorithm and an initial evaluation of the generated reports.

1 Introduction

Applications of Natural Language Generation (NLG) in the medical domain have been manifold. A new area where NLG could contribute to the improvement of services and to patient safety is pre-hospital care: care delivered to a patient before arrival at hospital. There are many challenges in delivering pre-hospital care, making it different from care taking place in the controlled circumstances of emergency departments or hospital wards.

Some Ambulance Services have developed innovative models to care for patients whilst an ambulance is en-route. Community First Responder (CFR) schemes recruit volunteers from local communities and give them the necessary training and equipment to deal with a limited range of medical emergencies. The premise is that even those with basic first-aid skills can save a life. It is their task to attend the casualty while waiting for the ambulance and to record their observations and actions on a paper patient report form (PRF). They may also assess the patient's physiological measurements (e.g. heart rate). In practice, due to time constraints, a verbal handover is performed and the PRF is filled in later. Physiological measurements may be written in ink on the back of a protective glove, and are rarely passed on in any systematic way.

The MIME (Managing Information in Medical Emergencies)[1] project is developing technology to support CFRs in the UK when they respond to patients. The project aims to enable CFRs to capture a greater volume of physiological patient data, giving them a better awareness of a patient's medical status so they can deliver more effective care.

There are two parts to our work: the use of novel lightweight wireless medical sensors that are simple and quick to apply, and the use of novel software that takes these inherently complex sensor data, along with some other information inputted by the user (e.g. patient demographics or actions performed) on a tablet computer, and present it very simply. We are working with two sensors that provide measurements of the patient's respiratory rate, heart rate and blood oxygen saturation. Our software can use NLG to produce a textual handover report at any time. This can be passed to an arriving paramedic to give a quick summary of the situation and can accompany the patient to inform later stages of care. We anticipate that our system will also provide some basic decision support based upon the patients clinical condition.

2 Related Work

Many situations arise in the medical domain where vast amounts of data are produced and their correct interpretation is crucial to the lives of patients. Interpreting these data is usually a demanding and complex task. Medical data are therefore often presented graphically or preferably in textual summaries (Law et al., 2005) making NLG important for various applications in the medical domain.

A number of systems address the problem of presenting medical information to patients in a form that they will understand. Examples are

[1] www.dotrural.ac.uk/mime

Proceedings of the 14th European Workshop on Natural Language Generation, pages 152–156,
Sofia, Bulgaria, August 8-9 2013.

STOP (Reiter et al., 2003), PILLS (Bouayad-Agha et al., 2002), MIGRANE (Buchanan et al., 1992), and Healthdoc (Hirst et al., 1997). Other systems, such as TOPAZ (Kahn et al., 1991) and Suregen (Hüske-Kraus, 2003), aim to summarise information in order to support medical decision-making.

In the case of MIME, the challenge is to summarise large amounts of sensor data, in the context of carer observations and actions, in a coherent way that supports quick decision making by the reader. The problem of describing the data relates to previous work on summarising time series data (e.g. (Yu et al., 2007)). In many ways, though, our problem is most similar to that of Babytalk BT-Nurse system (Hunter et al., 2012), which generates shift handover reports for nurses in a neonatal intensive care unit. The nature of the recipient is, however, different. Whereas BabyTalk addresses clinical staff in a controlled environment, MIME is aimed at people with little training who may have to deal with emergency situations very quickly. Further, while BT-Nurse works with an existing clinical record system, which does not always record all actions and observations which ideally would be included in a report, in MIME users enter exactly the information which MIME needs. This simplifies the NLG task, at the cost of adding a new task (interface construction).

3 The MIME project

In the first stage of MIME, we have developed a desktop application to prototype the generation of handover reports. We used simulated scenarios, where a panel of medical experts determined the sequence of events and predicted the stream of data from the simulated sensors.

The generated reports must provide a quick overview of the situation but at the same time be sufficiently comprehensive, while the format must enhance the readability. A general **structure for the handover reports** was determined in a user-centred development process together with ambulance clinicians. After the demographic description of the casualty and incident details (entered by the responder whenever they have an opportunity), two sections of generated text follow: the initial assessment section and the treatments and findings section. The initial assessment contains information on the patient gathered by the CFRs just after the sensors are applied and also any observations made during the first minute after the application of the sensors. The treatment and findings section is a report on the observations and actions of the CFRs while they waited for the ambulance to arrive. This includes a paragraph that sums up the condition of the patient at the time of handover.

Using sensors to capture physiological data continuously introduces the problem that irrelevant information needs to be suppressed in order not to overload the ambulance clinicians and hinder interpretation. The **NLG algorithm** that generates short as well as comprehensive handover reports accomplishes text planning in the two stages of document planning and micro-planning (Reiter and Dale, 2000). Document planning is responsible for the selection of the information that will be mentioned in the generated report. Events that will be mentioned in the text are selected and structured into a list of trees (similar to trees in Rhetorical Structure Theory (Scott and Sieckenius de Souza, 1990)). In the micro-planning step the structure of the document plan is linearised and sentences are compiled using coordination and aggregation.

Whereas some parts of the handover document (e.g. patient demographics) are relatively stylised, the main technically demanding part of the NLG involves the description of the "treatment and findings", which describes the events that happen whilst the patient is being cared for and relevant parts of the sensor data (see Figure 1). For this section of the report, the document planning algorithm is based on that of (Portet et al., 2007), which identifies a number of key events and creates a paragraph for each key event. Events that are explicitly linked to the key event or events that happen at the same time are added to the relevant paragraph. This is based on the earlier work of (Hallett et al., 2006).

4 Evaluation

In an initial evaluation we sought to assess how our reports would be received in comparison with the current situation – either short verbal reports or paper report forms (PRFs)– and also in comparison with what might be regarded as a "gold standard" report produced by an expert.

Materials: Two videos were produced independently of the NLG team, based on two scenarios of medical incidents typical of a CFRs caseload. These scenarios, a farm injury and chest pain, included a short description of the incident, similar

```
At 02:12, after RR remained fairly
constant around 30 bpm for 4 minutes,
high flow oxygen was applied, she took
her inhaler and RR decreased to 27
bpm. However, subsequently RR once
more remained fairly constant around
30 bpm for 8 minutes.

At 02:15 she was feeling faint.

At 02:15 the casualty was moved.

At 02:17 the casualty was once more
moved.
```

Figure 1: Part of the "Treatment and Findings" for an asthma scenario.

to the initial information a CFR would receive, a time line of events that happened before the ambulance arrived as well as simulated sensor data from the patient. The videos showed an actor in the role of CFR and another as patient, with the scenario time displayed in one corner. When the CFR performed readings of the physiological measures they were shown as subtitles.

The videos were presented to two CFRs and a paramedic, who were asked to imagine themselves in the situation of the CFR in the video, and to produce a handover report. Each video was only played once in order to produce more realistic results. We asked one CFR to construct a written "verbal" handover for the first scenario and to fill out a PRF for the other scenario, and the other CFR to do the "verbal" handover for the second scenario and to fill out the PRF for the first. To anonymise the PRF it was transcribed into a digital version. The paramedic received a blank sheet of paper and was requested to produce a handover report that he would like to receive from a CFR when arriving at the scene. Based on the scenarios we also generated two reports with the MIME system. This process resulted in four reports for each of the two scenarios, one transcribed verbal handover and a PRF from a CFR, a written handover report from a paramedic and the generated report.

Hypotheses: Our hypothesis was that the generated reports would improve on the current practice of verbal handovers and PRFs, and that paramedics would perceive them to be more suitable, hence rank them higher than the CFRs' verbal or PRF reports. The paramedic handover report might be regarded as a gold standard produced by an expert and we were interested in how the generated reports fared in comparison. Further, we hoped to gain information on how to improve our generated reports.

Participants: We approached paramedics in the Scottish Ambulance Service to participate in our study. Nine paramedics responded (eight male and one female; age range 32–56 years with 10–24 years' service).

Procedure: Participants received an invitation email with a link to a brief online survey and the eight reports as attachments. After an introduction and consent form they were forwarded to one of the two scenario descriptions and asked to rank the respective four reports. After that the participant was asked to rate the accuracy, understandability and usefulness of the generated report for this scenario on a 5-point Likert scale ranging from *very good* to *very bad* and to indicate what they liked or disliked about it in a free text box. This process was repeated for the second scenario.

4.1 Results

Ranking: An overview of the rankings can be found in Table 1. Apart from the rankings of participant 7 and 8, no large differences in how the reports were ranked could be observed between the two scenarios. We performed a Friedman test (Friedman, 1937) (farm injury scenario: chi-squared=4.3, df=3, p=0.23; chest pain scenario: chi-squared=12.44, df=3, p=0.006): some reports were ranked consistently higher or lower than others. The verbal CFR report was ranked worst in all but five cases. There is a high disparity in the rankings for the PRF, which was ranked first on eight occasions and in the other ten instances in third or fourth place. The generated report was ranked in first place only once, but eleven times in second place and in third place the other six times. In general the paramedic report, which was regarded as the "gold standard", was ranked better than the generated report, but in five cases the generated report was ranked better.

Rating: An overview of the ratings for the generated reports can be found in Table 2. The ratings for both scenarios were good on average, with a majority of ratings lying between very good to moderate. Only one rating (the accuracy of the generated report for the farm injury scenario) was bad; none was very bad. The ratings for the generated report of the chest pain scenario were on average better than those for the farm injury scenario. Accuracy had better ratings than usefulness and understandability in both scenarios.

Participant:	1	2	3	4	5	6	7	8	9	med	min	max
farm injury scenario												
Paramedic	2	2	3	1	1	3	3	2	1	2	1	3
Generated	3	3	2	2	2	2	2	3	2	2	2	3
CFR PRF	1	1	1	3	4	1	4	4	3	3	1	4
CFR verbal	4	4	4	4	3	4	1	1	4	4	1	4
chest pain scenario												
Paramedic	2	2	3	1	1	2	2	1	1	2	1	3
Generated	3	3	2	2	2	3	1	2	2	2	1	3
CFR PRF	1	1	1	3	4	1	4	3	3	3	1	4
CFR verbal	4	4	4	4	3	4	3	4	4	4	3	4

Table 1: Overview of the ranking results (*most preferred* (1) to *least preferred* (4)), median (med), maximum (max) and minimum (min) values for the patient report form (CFR PRF), paramedic report (Paramedic), generated report (generated) and verbal report (verbal CFR).

Participant:	1	2	3	4	5	6	7	8	9	med	min	max
farm injury scenario												
accuracy	1	2	1	4	2	2	1	1	1	1	1	4
useful.	3	3	2	2	1	2	2	1	1	2	1	3
unders.	2	3	2	2	1	3	3	1	1	2	1	3
chest pain scenario												
accuracy	2	2	1	1	1	3	1	2	1	1	1	3
useful.	2	3	2	1	1	2	1	1	1	1	2	3
unders.	2	3	2	1	1	3	2	1	1	2	1	3

Table 2: Overview of the rating results, median (med), maximum (max) and minimum (min) values for accuracy, usefulness (useful.) and understandability (unders.) of the generated reports, on a Likert scale (*very good* (1) to *very bad* (5)).

4.2 Discussion

We hypothesised that the generated reports would fare better than the verbal handovers and the PRFs. Results confirm a preference for the generated reports over the verbal handover. The paramedic reports, which were regarded as our "gold standard" were ranked higher than the generated reports. Interestingly, in almost half the cases there was a clear preference for the PRF and in the other cases the PRF ranked badly. This may have been affected by the familiarity of this medium and perhaps by the background assumption that this is how handover reports "should" be presented.

We regard this as a tentative confirmation that the generated texts compete favourably with the *status quo*. In a real world scenario the paramedics often get a verbal handover instead of the PRF and it should be noted that the PRF was printed and not handwritten. Furthermore, although the CFRs and paramedics only saw the scenario video once they were under no time pressure to submit the reports. Hence the quality of all the human reports in our experiment is likely to be better than normal.

Although each individual generally provided consistent responses across the two scenarios, there were variations between individuals. These different preferences may be merely stylistic choices or they may reflect in task performance. Preferences are not necessarily an indication of usefulness for a task (cf. (Law et al., 2005)).

In general the accuracy, understandability and usefulness of the generated reports received good ratings. Although participation was low, the qualitative data we gathered were valuable, every participant offered comments in the free text box on what they liked or disliked about the generated report. In general there seemed to be an impression that some sections were longer than necessary. One participant observed that reporting on observations a long time later is only useful if things have changed significantly. The structure and organisation of the report received some positive comments. For example one participant stated that he liked "the separate sections for information" and another commented that the report was "logically laid out", that it was "easy to obtain information" from the report and that it "clearly states intervention and outcome of intervention".

5 Conclusion and Future Work

Despite the fact that the experiment reported here involved a small number of participants, which implies that its results need to be interpreted with some caution, the generated reports produced by the MIME system appear to improve on the current practice of verbal handover. We aim to collect more responses and repeat the evaluation that has been presented. Our next step in evaluating the report generator will be to carry out a task based evaluation to see whether the preference ratings we have gathered can be reflected in performance measures.

We are now moving into the second stage of MIME and have started developing a new prototype, a mobile device that gets signals from two lightweight sensors. Here we will collect data from real emergency ambulance callouts by having a researcher join ambulance crews for their normal activity, which will be used to modify the NLG system (e.g. in order to allow for more reliable handling of noise).

6 Acknowledgments

This work is supported by the RCUK dot.rural Digital Economy Research Hub, University of Aberdeen (Grant reference: EP/G066051/1)

References

N. Bouayad-Agha, R. Power, D. Scott, and A. Belz. 2002. PILLS: Multilingual generation of medical information documents with overlapping content. In *Proceedings of LREC 2002*, pages 2111–2114.

B. Buchanan, J. Moore, D. Forsythe, G. Banks, and S. Ohlsson. 1992. Involving patients in health care: explanation in the clinical setting. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 510–514, January.

M. Friedman. 1937. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 32(200):675–701.

C. Hallett, R. Power, and D. Scott. 2006. Summarisation and visualisation of e-Health data repositories Conference Item Repositories. In *UK E-Science All-Hands Meeting*, pages 18–21.

G. Hirst, C. DiMarco, E. Hovy, and K. Parsons. 1997. Authoring and Generating Health-Education Documents That Are Tailored to the Needs of the Individual Patient. In Anthony Jameson, Cécile Paris, and Carlo Tasso, editors, *User Modeling: Proceedings of the Sixth International Conference, UM97*, pages 107–118. Springer Wien New York.

J. Hunter, Y. Freer, A. Gatt, E. Reiter, S. Sripada, and C Sykes. 2012. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial Intelligence in Medicine*, 56:157–172.

D. Hüske-Kraus. 2003. Suregen-2: A Shell System for the Generation of Clinical Documents. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2003)*, pages 215–218.

M. Kahn, L. Fagan, and L. Sheiner. 1991. Combining physiologic models and symbolic methods to interpret time-varying patient data. *Methods of information in medicine*, 30(3):167–78, August.

A. Law, Y. Freer, J. Hunter, R. Logie, N. McIntosh, and J. Quinn. 2005. A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. *Journal of clinical monitoring and computing*, 19(3):183–94, June.

F. Portet, E. Reiter, J. Hunter, and S. Sripada. 2007. Automatic generation of textual summaries from neonatal intensive care data. In *In Proccedings of the 11th Conference on Artificial Intelligence in Medicine (AIME 07). LNCS*, pages 227–236.

E. Reiter and R. Dale. 2000. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press.

E. Reiter, R. Robertson, and L. Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1-2):41–58, March.

D. Scott and C. Sieckenius de Souza. 1990. Getting the message across in rst-based text generation. In R. Dale, C. Mellish, and M. Zock, editors, *Current Research in Natural Language Generation*. Academic Press.

J. Yu, E. Reiter, J. Hunter, and C. Mellish. 2007. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13(1):25–49.