Working with Clinicians to Improve a Patient-Information NLG System

Saad Mahamood and Ehud Reiter
Department of Computing Science
University of Aberdeen
Aberdeen, Scotland, United Kingdom
{s.mahamood, e.reiter}@abdn.ac.uk

Abstract

NLG developers must work closely with domain experts in order to build good NLG systems, but relatively little has been published about this process. In this paper, we describe how NLG developers worked with clinicians (nurses) to improve an NLG system which generates information for parents of babies in a neonatal intensive care unit, using a structured revision process. We believe that such a process can significantly enhance the quality of many NLG systems, in medicine and elsewhere.

1 Introduction

Like other artificial intelligence (AI) systems, most Natural Language Generation (NLG) systems incorporate domain knowledge (and domain communication knowledge (Kittredge et al., 1991)), either implicitly or explicitly. Developers must work with domain experts to acquire such knowledge. Also like software systems in general, applied NLG systems must meet domain and application specific requirements in order to be useful; these again must come from domain experts.

Since very few domain experts are familiar with NLG, it is usually extremely difficult to acquire a complete set of requirements, domain knowledge, and domain communication knowledge at the beginning of an NLG project. Especially, if no pre-existing "golden standard" corpus of domain texts exists. Indeed, in many cases domain experts may find it difficult to give detailed requirements and knowledge until they can see a version of the NLG system working on concrete examples. This suggests that an iterative software development methodology should be used, where domain experts repeatedly try out an NLG system, revise underlying domain (communication) knowledge and request changes to the system's functionality, and wait for developers to implement these changes before repeating the process.

We describe how we carried out this process on BabyTalk-Family (Mahamood and Reiter, 2011), an NLG system which generates summaries of clinical data about a baby in a neonatal intensive care unit (NICU), for the babys parents. Over a 6 month period, this process enabled us to improve an initial version of the system (essentially the result of a PhD) to the point where the system was good enough to be deployable live in a hospital context. We also describe how the feedback from the clinicians changed over the course of this period.

2 Previous Research

Reiter et al. (2003) describe a knowledge acquisition strategy for building NLG systems which includes 4 stages: *directly asking domain experts for knowledge*, *structured knowledge acquisition activities with experts*, *corpus analysis*, and *revision with experts*. In this paper we focus on the last of these phases, revision with experts. Reiter et al. describe this process in high-level qualitative terms; in this paper our goal is to give a more detailed description of the methodology, and also concrete data about the comments received, and how they changed over time.

The most similar previous work which we are

INLG 2012 Proceedings of the 7th International Natural Language Generation Conference, pages 100–104,
Utica, May 2012.

aware of is Williams and Reiter (2005), who describe a methodology for acquiring content selection rules from domain experts, which is also based on an iterative refinement process with domain experts. Their process is broadly similar to what we describe in this paper, but they focus just on content selection, and do not give quantitative data about the revision process.

In the wider software engineering community, there has been a move to iterative development methodologies, instead of the classic "waterfall" pipeline. In particular, agile methodologies (Martin, 2002) are based on rapid iterations and frequent feedback from users; we are in a sense trying to apply some ideas from agile software engineering to the task of building NLG systems. Our methodology also can be considered to be a type of user-centred design (Norman and Draper, 1986).

3 BabyTalk-Family

BabyTalk-Family (Mahamood and Reiter, 2011) generates summaries of clinical data about babies in a neonatal intensive care unit (NICU) for parents. For more details about BabyTalk-Family, including example outputs, please see Mahamood and Reiter.

BabyTalk-Family (BT-Family) was initially developed as part of a PhD project (Mahamood, 2010). As such it was evaluated by showing output texts (based on real NICU data) to people who had previously had a baby in NICU; the texts did not describe the subject's own baby (i.e., the subjects read texts which summarised other people's babies; they had no previous knowledge of these babies). BT-Family was also not rigorously tested from a software quality assurance perspective. The work presented here arose from a followup project whose goal was to deploy BT-Family live in a NICU, where parents who currently had babies in NICU could read summaries of their baby's clinical data. Such a deployment required generated texts to be of much higher quality (in terms of both content and language); we achieved this quality using the revision process described in this paper.

BT-Family is part of the BabyTalk family of systems (Gatt et al., 2009). All BabyTalk systems use the same input data (NICU patient record), but they produce different texts from this data; in particular BT45 (Portet et al., 2009) produces texts which summarise short periods to help real-time decision making by clinicians, and BT-Nurse (Hunter et al., 2011) produces summaries of 12 hours of data for nurses, to support shift handover. BT-Nurse was also deployed in the ward, to facilitate evaluation by nurses who read reports about babies they were currently looking after. To support this deployment, the BT-Nurse developers spent about one month carrying out a revision process with clinicians, in a somewhat unstructured fashion. One outcome of the BT-Nurse evaluation was that the system suffered because the revision process was neither sufficiently well structured nor long enough; this was one of the motivations for the work presented here.

4 Revision Methodology

The revision process was carried out at the Neonatal Intensive Care Unit in conjunction with the hospital Principal Investigator (PI) of our project and two research nurses. We started with an initial familiarisation period for the nurses (the hospital PI was already familiar with BT-Family), where we explained the goals of the project and asked the nurses to examine some example BT-Family texts, which we then discussed.

After the nurses were familiar with the project, we conducted a number of revision cycles. Each cycle followed the following procedure:

1. The clinicians (either the hospital PI or the research nurses) choose between 3 and 11 scenarios (one day's worth of data from one baby). These scenarios were chosen to test the system against a diverse range of babies in different clinical conditions; scenarios were also chosen to check whether issues identified in previous cycles had been addressed.

2. The nurses examined the texts generated by BT-Family for the chosen scenarios. They both directly commented on the texts (by writing notes on hard-copy), and also (in some cases) edited the texts to show what they would have liked to see.

3. The NLG developers analysed the comments and revised texts; distilled from these a list of specific change requests; prioritised the change requests on the basis of importance and difficulty; and implemented as many change requests as possible given the time constraints of the cycle.

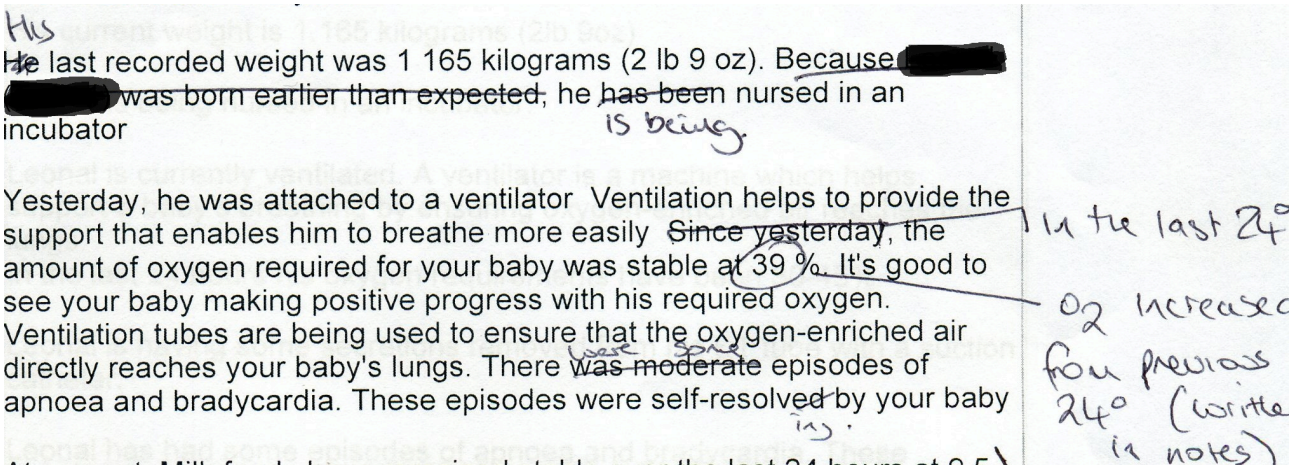

Figure 1: Example of marked up text annotated by a research nurse. The baby's forename has been blacked out.

4. The scenarios were rerun through the updated system, and the NLG developers checked that the issues had been addressed. Clinicians did not usually look at the revised texts, instead they would check that the issues had been resolved in new scenarios in the next cycle.

The above process was carried out 14 times over a 6 month period with each cycle taking on average 11.28 days. A research fellow (Saad Mahamood) was assigned to implement these changes working full-time over this 6 month period. The length between each revision cycle was variable due to the availability of the domain experts and the variable level of complexity to implement identified changes to the BT-Family system.

Figure 1 shows a extract from an early BT-Family text generated in July 2011 that needed a lot of revision. In this example, the nurse has identified the following issues:

- Incorrect pronoun: *He* instead of *His*.
- Unnecessary phrase: *Because XXXX was born earlier than expected.*
- Change in tense: *is being* instead of *has been*.
- Change in wording of time phrase: *In the last 24 hours* instead of *Since yesterday*.
- Incorrect content: incubator oxygen has increased, it is not stable.
- Grammar mistake: *were* instead of *was*.
- Change in content: *some* (frequency) instead of *moderate* (severity).
- Change in wording: *self-resolving* instead of *self-resolved*.

5 Analysis of Feedback over Time

We extracted hand-written comments on BT-Family texts (of the type shown in Figure 1) and annotated the comments using a scheme similar to that used by Hunter et al (2011) for analysing comments on BT-Nurse texts. Two annotators were used with the first annotating the entire set of 75 reports using a pre-agreed classification scheme. The classification scheme that was used consisted of three types of categories: *Content Errors*, *Language Errors*, and *Comments* with each containing specific categorisation labels as shown in Table 1. Content Errors labels were used to annotate comments when there were content based mistakes. Language error labels were used to categorise the different types of language based mistakes. Finally, comment labels were used to classify different types of comments made by the nurses. The second annotator annotated a random partial subset of the reports independently to check for the level of agreement between the first and second annotators. By using Cohen's kappa coefficient we found the level of inter-annotator agreement was *k=0.702*.

Content errors were the most predominate type of annotation (*50.54%*), followed by Language errors (*25.18%*), and comments (*24.27%*). Positive comments were unusual (only 5 in total), because the clinicians were explicitly asked to focus on prob-

Content Errors	Language Errors	Comment
unnecessary *(44.20%)*	spelling mistake *(8.14%)*	positive *(3.75%)*
missing *(28.26%)*	grammar mistake *(22.22%)*	negative *(0.75%)*
wrong *(22.82%)*	incorrect tense/aspect *(18.51%)*	no agreement *(1.50%)*
should-be-elsewhere *(4.71%)*	different word(s) required *(35.55%)*	reformulation *(12.78%)*
	unnecessary words *(3.70%)*	observation *(66.16%)*
	precision/vagueness *(11.85%)*	question *(15.03%)*

Table 1: List of annotation categories and the labels within each category that was used. The frequency for each label in it's category is given in brackets.

Month	Number of revision cycles	Avg. scenarios per cycle	Avg. number of content errors	Avg. number of language errors	Avg. number of comments
June	1	5	1.8	4.2	1.2
July	2	8	4.93	5.5	1.87
August	2	5	4.8	4	5.8
September	2	4	6.37	8.5	4
October	3	7	2.95	1.57	6.42
November	3	5	1.6	1.6	3.6
December	1	5	0.8	0	0.4
Overall	14	5.7	6.92	3.62	3.32

Table 2: Summary table showing the average number of content errors, language errors, and comments per scenario.

lems. Table 2 shows statistics for the revision process per month; the process started in the second half of June, and ended in the first half of December.

From a qualitative perspective, the data suggests that there were two phases to the revision process. In the first phase (June to September), the number of content and language errors in fact went up. We believe this is because during this phase we were adding around 16 new types of content to the reports (based on requests from the clinicians) as well as fixing problems with existing content (of the sort shown in Figure 1); this additional content itself often needed to be revised in subsequent revision cycles, which increased the error count for these cycles. These additional errors from the addition of new content may of arisen due to the complexity and variation of clinical data. Additionally, our 3-year old anonymised test set of clinical data may not of been as representative as the live data due to changes/additions in patient data. In the second phase (October to December), requests for new content diminished (around 4 requests) and we focused on fixing problems with existing content; in this phase, the number of content and language errors steadily decreased (that is, the system improved from the clinician's perspective), until we reached the point in mid December when the clinicians were satisfied that the quality of BT-Family texts was consistently good from their perspective.

When the revision process ended, we started evaluating BT-Family texts directly with parents, by showing parents texts about their babies. This work is ongoing, but initial pilot results to date indicate that parents are very happy with the texts, and do not see major problems with either the language or the content of the texts.

6 Discussion

The revision process had a major impact on the quality of BT-Family texts, as perceived by the clinicians. At the start of the process (June 2011), the texts had so many mistakes that they were unusable; the clinicians would not allow us to show parents BT-Family texts about their babies, even in the context of a pilot study. After 14 revision rounds over a 6 month period, text quality had improved dramatically, to the point where clinicians allowed us to start working directly with parents to get their feedback and comments on BT-Family texts.

The fact that a new set of scenarios was used in every iteration of the revision process was essen-

tial to giving clinicians confidence that text quality would be acceptable in new cases; they would not have had such confidence if we had focused on improving the same set of texts.

The revision process took 6 months, which is a considerable amount of time. This process would have been shorter if BT-Family had undergone a more rigorous testing and quality assurance (QA) process ahead of time, which would for example have addressed grammar mistakes, and (more importantly) tested the system's handling of boundary and unusual cases. The process probably could also have been further shortened in other ways, for example by performing 3 revision cycles per month instead of 2.

However, one reason the process took so long was that the functionality of the system changed; as the clinicians got a better idea of what BT-Family could do and how it could help parents, they requested new features, which we tried to add to the system whenever possible. We also had to accommodate changes in the input data (patient record), which reflected changes in NICU procedures due to new drugs, equipment, procedures, etc. So we were not just tweaking the system to make it work better, we were also enhancing its functionality and adapting it to changing input data, which is a time consuming process.

7 Conclusion

We have presented a methodology for improving the quality and appropriateness of texts produced by applied NLG systems, by repeatedly revising texts based on feedback from domain experts. As we have show in the results, the process is a time consuming one, but appears to be quite effective in bringing an NLG system to the required level of quality in a clinical domain.

Acknowledgements

This work is funded by the UK Engineering and Physical Sciences Council (EPSRC) and Digital Economy grant EP/H042938/1. Many thanks to Dr. Yvonne Freer, Alison Young, and Joanne McCormick of the Neonatal Intensive Care Unit at Simpson Centre for Reproductive Health, Royal Infirmary of Edinburgh Hospital, for their help.

References

Albert Gatt, Francois Portet, Ehud Reiter, Jum Hunter, Saad Mahamood, Wendy Moncur, and Somayajulu Sripada. 2009. From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. *AI Communications*, 22(3):153–186.

James Hunter, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, Cindy Sykes, and Dave Westwater. 2011. BT-Nurse: Computer generation of natural language shift summaries from complex heterogeneous medical data. *Journal of the Americal Medical Informatics Association*, 18(5):621–624.

Richard Kittredge, Tanya Korelsky, and Owen Rambow. 1991. On the need for domain communication language. *Computational Intelligence*, 7(4):305–314.

Saad Mahamood and Ehud Reiter. 2011. Generating affective natural language for parents of neonatal infants. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 12–21, Nancy, France, September. Association for Computational Linguistics.

Saad Mahamood. 2010. *Generating Affective Natural Language for Parents of Neonatal Infants*. Ph.D. thesis, University of Aberdeen, Department of Computing Science.

Richard Martin. 2002. *Agile Software Development, Principles, Patterns, and Practices*.

Donald A. Norman and Stephen W. Draper. 1986. *User Centered System Design; New Perspectives on Human-Computer Interaction*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816.

Ehud Reiter, Somayajulu Sripada, and Roma Robertson. 2003. Acquiring correct knowledge for natural language generation. *Journal of Artificial Intelligence Research*, 18:491–516.

Sandra Williams and Ehud Reiter. 2005. Deriving content selection rules from a corpus of non-naturally occurring documents for a novel NLG application. In *Proceedings of Corpus Linguistics workshop on using Corpora for NLG*.