Text content and task performance in the evaluation of a Natural Language Generation system

Albert Gatt
Department of Computing Science
University of Aberdeen
a.gatt@abdn.ac.uk

François Portet
Laboratoire d'Informatique de Grenoble
Grenoble Institute of Technology
francois.portet@imag.fr

Abstract

An important question in the evaluation of Natural Language Generation systems concerns the relationship between textual characteristics and task performance. If the results of task-based evaluation can be correlated to properties of the text, there are better prospects for improving the system. The present paper investigates this relationship by focusing on the outcomes of a task-based evaluation of a system that generates summaries of patient data, attempting to correlate these with the results of an analysis of the system's texts, compared to a set of gold standard human-authored summaries.

Keywords

Natural Language Generation, evaluation, decision support, domain ontology

1 Introduction

In the evaluation of NLP systems, an important distinction is that between *intrinsic* criteria, which typically measure aspects of output quality, and *extrinsic* ones, which assess a system in terms of its adequacy in its target setting [9]. The relationship between the two is important. A systematic relationship between the outcomes of extrinsic evaluation and properties of a system's output can indicate directions for improvement in output, leading to improvements in the system's utility in its target setting.

This paper focusses on the evaluation of a a Natural Language Generation (NLG) system, BT-45 [13], which generates summaries of clinical patient data in a Neonatal Intensive Care Unit (NICU). It was been evaluated in an experiment comparing decision making by clinicians based on the system output and other ways of presenting the same information, including human-authored summaries [17]. The aims of this paper are twofold. First, we investigate the relationship between intrinsic properties of generated text – particularly its informativeness and relevance – and its utility for decision-making. Second, we propose a novel intrinsic evaluation method. Rather than comparing BT-45 texts to a gold standard based on surface characteristics, such as matching $n-$grams, we make explicit use of domain knowledge in the form of an ontology and attempt to quantify the differences in the content selection strategies underlying the BT-45 and the gold standard texts.

2 Background

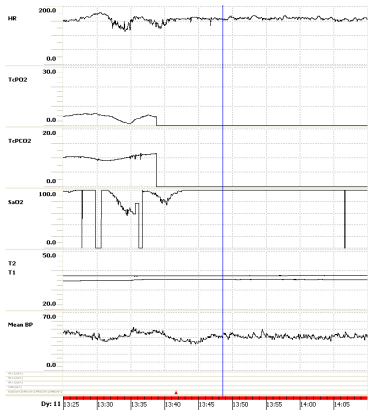
Intrinsic evaluation in NLG has often relied on human input, typically in the form of ratings of or responses to questionnaires [12, 4, 7]. Automatic intrinsic methods exploiting corpora have mostly been used in evaluations of morphosyntactic realisers [11, 3]. Extrinsic, task-based methods are also widespread [14, 10, 15]. While such studies tend to be more expensive and labour-intensive, extrinsic evaluation criteria give a reliable assessment of the system's utility in doing what it was designed to do.

The relationship between these different classes of evaluation methods is not straightforward. Recent work has shown that corpus-based intrinsic methods do not correlate with the results of intrinsic evaluation based on human judgements, suggesting that they are measuring different aspects of output quality [2]. Cautionary notes have also been sounded in Machine Translation [5] and summarisation [6], with some recent work in NLG showing that intrinsic measures also correlate poorly with task-based measures NLG [1]. On the other hand, the relationship between task-based measures and textual characteristics bears on several important questions. Task-based evaluations tend to yield global scores from which it is often hard to extract specific indicators of a system's weaknesses. While judgement elicitation studies may be better suited to this purpose, these do not necessarily reflect a system's utility in a task, while corpus-based studies necessarily depend on a finite number of reference outputs against which to compare a system, which do not exhaust the space of possibilities.

3 The BT-45 system and evaluation

In this section, we briefly summarise the main aspects of the BT-45 architecture and the task-based evaluation (see [13] and [17] for a complete description). BT-45 produces a textual summary of 45 minutes of NICU data, in the form of physiological signals measured from a patient using sensors, and data relating to discrete events, which are logged by clinicians in a database in the course of a shift. A snapshot of the input data is displayed in Figure 1(a). Figure 1(b) shows a summary of this data written by two expert neonatologists, while Figure 1(c) presents the BT-45 output for the same period. As the summary shows, BT-45 texts kept interpretation and diagnosis to a minimum, generating a descriptive summary of the salient events related to a patient. This involved a four-stage process, each making use of a domain-specific ontology. First, the main features of the signal data are identified, and the discrete data are extracted. Both form the input to a *data interpretation* stage,

International Conference RANLP 2009 - Borovets, Bulgaria, pages 107–112



(a) Graphical presentation

You saw the baby between 13:25 and 14:10.
At the start of the monitoring period: HR baseline is 145-155, oxygen saturation is 99%, pO2 = 4.9 and CO2 = 10.3 Mean BP is 37-47; T1 and T2 are 37.3degC and 34.6degC respectively.
At 13:33 there is a desaturation to 59%, which is accompanied by a drop in pO2 to 1.3 and a decrease in HR to 122. The blood pressure rises toward the end of this episode to 49. These parameters return to their baselines by 13:37.
At 13:38 the transcutaneous probe is taken off the baby to be re-sited. With this there is a desaturation to 80% and a decrease in HR to 129. The FiO2 is changed to 27%. The BP decreases to 33, then rises over the next 2 minutes to a baseline of 42 until the end of the trace.
By 13:43 the HR = 155 and oxygen saturation = 100%. The baby's observations remain steady around these values until the end of the monitoring period.

(b) Human-authored summary

You saw the baby between 13:25 and 14:10. Heart Rate (HR) = 149. Core Temperature (T1) = 37.4. Peripheral Temperature (T2) = 34.6. Transcutaneous Oxygen (TcPO2) = 4.8. Transcutaneous CO2 (TcPCO2) = 10.2. Mean Blood Pressure (mean BP) = 45. Oxygen Saturation (SaO2) = 96.
Over the next 38 minutes T1 stayed at around 37.4.
By 13:33 TcPO2 had rapidly decreased to 2.7. Previously HR had increased to 173.
By 13:35 there had been 2 successive desaturations down to 56. Previously T2 had increased to 35.5.
By 13:40 SaO2 had rapidly decreased to 79.
Over the next 30 minutes HR gradually increased to 157.
At 13:41 Fraction of Inspired Oxygen (FIO2) was set to 27%. Previously mean BP had decreased to 35.

(c) BT-45 summary

Fig. 1: NICU *data presented in different formats*

which creates abstractions from the raw data and relates events to eachother via causal and other links. *Document planning* selects important events and structures them into a document plan, while *microplanning* fleshes out their semantic content and realises the text.

The system was evaluated during an off-ward experiment during which 35 clinicians in 4 groups (junior and senior doctors and junior and senior nurses) were asked to select the most appropriate clinical actions to take in relation to a patient, based on around 45 minutes of data. 24 scenarios were presented to each participant in one of the three conditions shown in Figure 1: **(G)** Graphically; **(H)** textual summary written by human experts; and **(C)** textual summary generated by BT-45. Participants were asked to select the appropriate clinical actions to take at the end of the period from a predefined set of 18 actions. Prior to the experiment, a senior neonatal nurse and consultant neonatologist identified, for each scenario, the subsets of appropriate, inappropriate (potentially harmful) and neutral actions from this set. Moreover, for each scenario there was one *target action* which was deemed to be the most important out of all the appropriate ones. For each participant p and scenario s, a performance score S_s^p was computed based on the proportion P_{AP_s} of actions selected out of the set of appropriate actions for the scenario, AP_s, and the proportion P_{INAP_s} selected out of the set of inappropriate actions INAP_s:

$$S_s^p = P_{\text{AP}_s} - P_{\text{INAP}_s} \in [-1, 1] \quad (1)$$

Overall, the human expert texts (H) led to the best decision making ($.45^{SD=.10}$) followed by the Computer texts (C) ($.41^{SD=.13}$) and the Graphical (G) condition ($.40^{SD=.15}$). A 3 (Condition) x 4 (Group) by-subjects ANOVA showed no main effect of participant Group, but an effect of Condition that approached significance ($F(2, 31) = 2.939, p = 0.06$). There was no interaction. Separate ANOVAs showed a difference between the G and H conditions ($F(1, 31) = 4.975, p < 0.05$) and the C and H conditions ($F(1, 31) = 5.266, p < 0.05$), but none between G and C. Thus, human texts proved most useful to decision-making, but generated texts were found to be no worse than presentation in the graphical condition, which is the modality used in current clinical practice. A follow-up analysis comparing the scenarios based on their main target action showed a significant difference in performance between H and C texts ($F(1, 7) = 8.002; p < 001$).

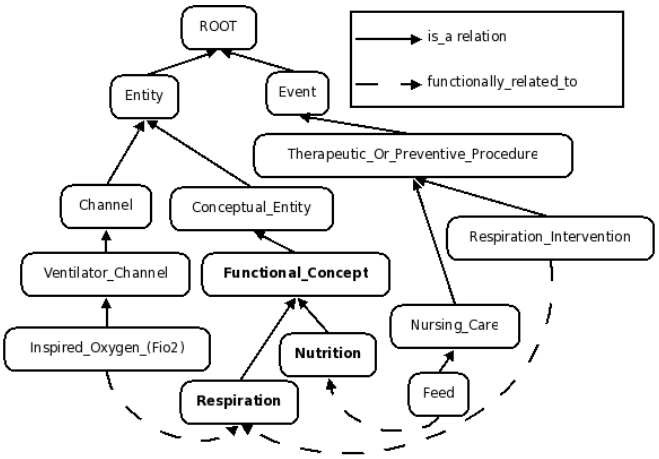


Fig. 2: *Excerpt from the ontology showing relations to functional concepts (in bold)*

Although this evaluation tells us that textual presentation can be very effective, it does not give direct information about which aspects of the experimental texts played a role in decision-making. On the other hand, the significant effect of main target action does suggest that some of the burden is carried by the content selection strategies in the H and C texts, with the human-authored summaries incorporating more information that was relevant to the appropriate actions. This was the focus of our follow-up evaluation.

4 The role of the ontology

Much of the processing in BT-45 relied on an ontology, an excerpt of which is shown in Figure 2. The complete ontology represented more than 550 concepts, though only a subset was used to a significant extent in BT-45. The principal top-level nodes are EVENT and ENTITY. The latter subsumes domain objects, such as VENTILATOR, which are not subject to significant change over time, whereas

EVENTs are inherently temporal and subsume INTERVENTION (e.g. drug administration), OBSERVATION (e.g. the observation that a baby has poor capillary refill), DATA COLLECTION (e.g. adjusting sensors), etc. While the ontology was a relatively static repository of declarative knowledge, procedural knowledge used for inference (e.g. causal associations between concepts, abstractions, etc), was encoded in production rules in the data interpretation module.

Following the BT-45 evaluation, a senior consultant was asked to udpate the ontology by linking entities and events to functional concepts (FC) corresponding to physiological systems, such as BLOOD PRODUCTION and RESPIRATION. These `functionally_related_to` links, shown in Figure 2, reflect the tendency of clinicians to link concepts in terms of their clinical significance (rather than, say, their structure). The underlying principle is that each event is related to an FC that clinicians monitor, and each clinical action has a purpose that is itself related to one or several FCs (e.g. VENTILATION to support RESPIRATION). A reasoning module can thus relate different events via their common relations to FCs. For example, if an alteration of the FiO2 (Fraction of inspired oxygen) on the ventilator is recorded, then this event can be linked to all other events that are related to RESPIRATION. In what follows, we shall use $\text{FC}(x)$ to abbreviate the set of functional concepts that a concept x is functionally related to.

5 Quantitative evaluation methods

To identify the characteristics of the texts which contributed to the decision-making performance, two kinds of quantitative methods were used. The first compared the computer (C) and human-authored (H) texts in terms of *informativeness*. This was based on the expectation that differences in informativeness between texts would covary with differences in decision-making, since more information should increase the likelihood of making the correct decisions. However, the extent to which information impacts decision-making also depends on the *relevance* of the information. Hence, the second evaluation measure quantified the extent to which a text was relevant to the appropriate actions on a given scenario. Presumably, higher relevance would give rise to a greater probability of a reader making the right decisions. Both methods relied on a prior annotation of the H and C texts, using the ontology as the central repository of domain knowledge.

Corpus annotation The corpus consists of 48 texts representing the 24 scenarios generated by BT-45 and the 24 scenarios written by the clinical experts. Annotation was consensus-based and was carried out manually by the authors, with a scheme to mark up text segments representing an event or an aggregation of events, as shown in the excerpt below.

```
<EVENT CARDINALITY="3" TYPE="TREND"
       SOURCE="TcPO2,HR,mean BP">
   These parameters return to their baselines
</EVENT>
[...]
<EVENT CARDINALITY="1" TYPE="RE-SITE_PROBES"
       DONE_TO="TCM_SENSOR">
   the transcutaneous probe is [...] re-sited
</EVENT>
```

An EVENT tag has a TYPE whose value is an ontology concept; CARDINALITY is used to account for reference to multiple events (e.g. *These parameters* in line 1, which refers to several physiological parameters mentioned earlier in the example text); optionally, the SOURCE attribute indicates the location of an event (usually a physiological parameter, such as heart rate in the case of trends like *a decrease in HR*), while the DONE_TO attribute is used when the event involves the use or modification of an entity (e.g. an instrument or a drug). Once the texts were marked up, the relation of an event to its FC was found by instantiating the event in the ontology based either on its TYPE attribute or the value of its DONE_TO or SOURCE attributes, to retrieve the value of its `functionally_related_to` property.

Informativeness Informativeness of a text was computed as a global estimate of the amount of information conveyed, irrespective of what the appropriate decisions to be taken were. This measure indicates whether the C and H texts convey a different amount of descriptive information, which could explain some differences in decision-making performance. In this work, we used two definitions of informativeness: (i) the number of (clinical) events NE that a text references; (ii) the length of the text in tokens (words) NW. Differentiating them allows us to distinguish between informativeness based solely on events (NE), and informativeness which also includes expressions which are not annotated (including adverbials and discourse connectives).

Relevance The relevance of a text for a given experimental scenario s was defined in terms of whether the events it mentioned have some relationship to the clinical actions which are appropriate for that scenario (AP_s). Let $\text{E}_{s,t}$ be the set of events mentioned in text t for scenario s. The relevance of an event is defined as follows:

Definition 1 (relevance) *An event* $e \in \text{E}_{s,t}$ *is relevant iff* $\exists a \in \text{AP}_s : \text{FC}(e) \cap \text{FC}(a) \neq \emptyset$

Of course an event can be relevant to more than one action. Likewise, we define the *ir*relevance of an event e if it is functionally associated to an element of the set of inappropriate actions for a scenario (INAP_s).

Though this definition gives a fair handle on the notion of relevance, it only approximates the information clinicians bring to bear on their decisions. Our hypothesis is that an appropriate action which is related to some FCs is more likely to be taken if these FCs are referenced in the text by mentioning events which are functionally related to them. For instance, if a text gives no information related to RESPIRATION, a clinician cannot make a decision related to managing a patient's artificial ventilation. Another noteworthy aspect of this method is its reliance on the knowledge (i.e. the ontology) that is already available in the system, rather than on human expert judgements, which could be subject to expert bias. For both human and computer texts, two scores were computed: $\text{REL}_{s,t}$, the number of relevant events in text t for scenario s; and $\text{IRREL}_{s,t}$, the number of irrelevant events in the text.

As defined above, relevance does not take into account the prior probability of the actions. In a clinical environment, some actions, such as taking a blood gas from an arterial line, are performed routinely, while others, such as

Table 1: *Examples of prior probability of actions*

Action	P(Action)
blood gas	0.0034
monitor equipment	0.0077
CPR	2.28E-5
manage temperature	0.0071
manage ventilation	0.0326
CPAP	0.0015

	Human (H)		BT-45 (C)		**Overall**
	Mean (SD)	*r*	*Mean (SD)*	*r*	r_{diff}
NW	146.75 (86.8)	-.24	122.33 (41.4)	-.47*	.434*
NE	19.75 (12.5)	-.25	17.35 (6.4)	-.47*	.458*
REL	0.326 (.37)	-.09	0.217 (.17)	-.10	.16
IRREL	0.324 (.31)	-.15	0.242 (.16)	-.32	.25

Table 2: *Mean scores across groups on informativeness and relevance, with correlations. r = Pearson's correlation between intrinsic and extrinsic performance score. r_{diff} = correlation between difference in performance and difference in score between H and C. Asterisk (∗) indicates significance at $p \leq .05$*

resuscitating a patient, are only taken in exceptional circumstances. To account for the potential effects of this on decision making, we computed the prior probability of each clinical action used in the experiment, as a maximum likelihood estimate based on a large database of 43,889 clinical actions recorded by an on-site research nurse over a period of 4 months in a NICU [8]. Some example probabilities for each action are displayed in Table 1. These were used to weight the relevance scores, which are now defined as follows:

$$\text{REL}_{s,t} = \sum_{a \in \text{AP}_s} P(a).N(a)_t \quad (2)$$

$$\text{IRREL}_{s,t} = \sum_{a \in \text{INAP}_s} P(a).N(a)_t \quad (3)$$

where a is any action and $N(a)_t$ stands for the number of times action a was referenced by the events in text t.

6 Evaluation results

In this section, we present the results of our comparison of the 24 pairs of human and computer-generated texts. All results are reported using scenarios ($N = 24$) as source of variance. For each measure, we report (a) the correlation (r) between the performance score and the measure in a given condition (H or C); (b) differences between H and C on the measure; (c) the correlation between the *absolute* differences in the scores obtained by H and C texts and those in decision-making performance (r_{diff}). Table 2 displays all descriptives and correlation coefficients. In what follows, we summarise the main observations, discussing their implications in Section 7.

6.1 Effect of informativeness

The means for NE and NW in Table 2 indicates that human texts tend to mention more events and are slightly longer than the BT-45 texts, but with much higher variability (higher SD) between scenarios. No significant correlations (r) were observed between NE or NW and performance in the H condition; in the C condition, both correlations are significant. However, in all cases, the r scores are negative, suggesting that more information tended to be linked to *lower* decision-making performance. Paired t-tests showed that there was no significant difference between the two sets of texts on NE ($t(23) = 1.24, p > .2$), though the difference on the NW score approached significance ($t(23) = 1.90, p = .07$).

There was a significant positive correlation between the absolute differences in informativeness scores and decision-making (r_{diff}). We also investigated the correlation within user groups, finding a significant correlation only in the case of Senior Doctors for both measures (NW: $r = 45$; NE: $r = .53$; $p \leq .05$). The differences in scores are plotted against differences in performance in Figures 3(a) and 3(b). In both cases, a linear relationship accounts for approximately 20% of the variance, as reflected by the R^2 value associated with the regression line. Omitting the 'outliers' where the differences between H and C in NW and NE seems highest (4 points in Figure 3(a), 5 in Figure 3(a)) does not improve the models.

In summary, human expert texts shown more variability than BT-45 texts on our informativeness measures. Separate correlations for H and C show a surprising negative covariation between the measures and decision-making, while a weak positive relationship can be observed between the difference in the scores and decision-making differences in the two conditions.

6.2 Effect of relevance

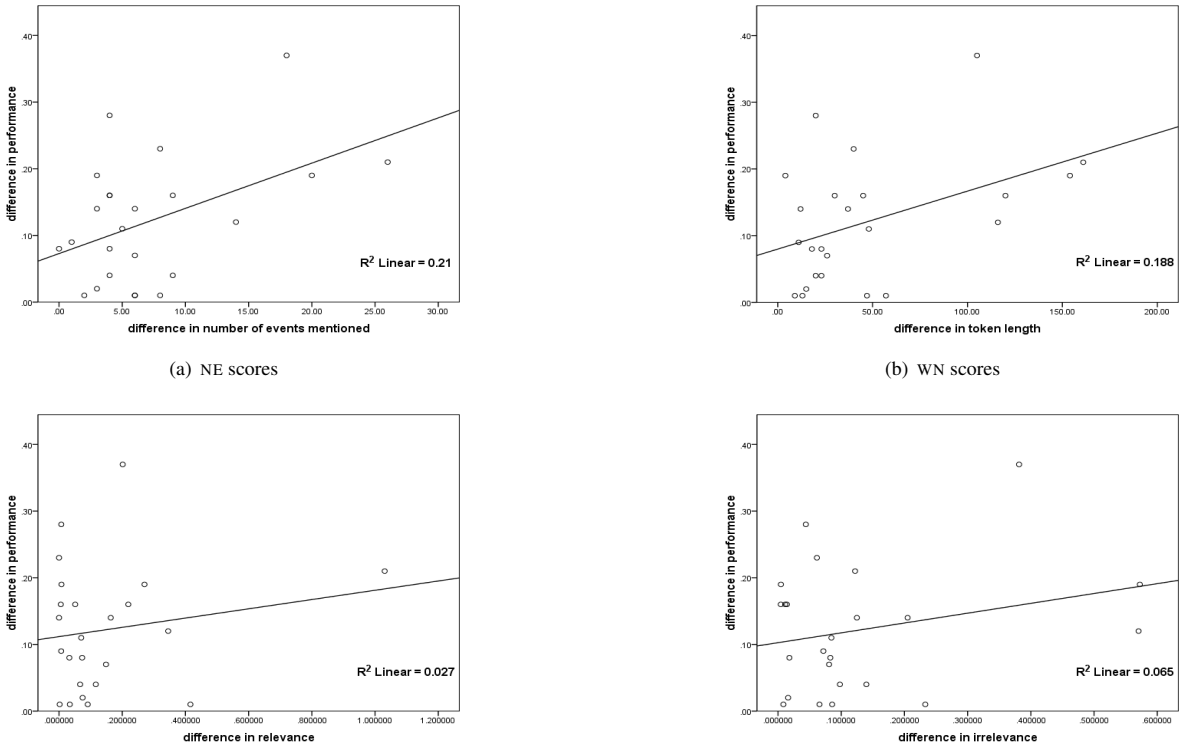
Table 2 shows that the H texts achieved higher scores on both REL and IRREL, that is, the events mentioned by human experts referenced more appropriate actions, but also more 'inappropriate' ones. Once again, the variation is higher in the H texts. There were significant differences on both measures between H and C texts (REL: $t(23) = 2.23, p < .05$; IRREL: $t(23) = 2.11, p < .05$). As for the correlations within each condition (r), they are once again negative, and never reach significance.

On the other hand, we do observe a significant positive correlation between difference in performance and both REL and IRREL. Once again, this seems to be primarily due to the Senior Doctor group, where differences in REL and difference in performance were significantly correlated ($r = .54; p < .05$). The correlations were not significant for any other user group, and never significant for differences in IRREL.

These trends are further reflected in Figures 3(c) and 3(d), which again plot the difference in performance between H and C against difference in the REL and IRREL. Once again, both figures show that the linear relationship accounts for a very low proportion of the variance ($< 2\%$); removing the outliers in either case again results in no improvement.

7 Discussion

The results reported above do not offer very strong support for our initial hypotheses that differences in informativeness or relevance would account for differences in performance. Despite some significant differences in content, as shown by the difference between H and C on REL and IRREL, the relationship between content and task performance is weak at best. Although this is consistent with previous results comparing intrinsic and extrinsic evaluation



(a) NE scores

(b) WN scores

(c) REL scores

(d) IRREL scores

Fig. 3: *Difference in task performance against differences in* NW*,* NE*,* REL *and* IRREL *scores*

measures, taking the present conclusions as final would seem to be premature, for a number of reasons which we outline in this section.

Content vs. structure In focussing exclusively on a definition of content based on the events referenced by a text, our evaluation scores ignore aspects of discourse and information structure, such as the extent to which the texts link the events mentioned via discourse or temporal connectives, as in the human-authored excerpt below:

Example 1 (Human text) *The pCO2 continues to rise to 10.1. The baby is pale and unresponsive. ET suction is given, baby is turned and at 17:02 the ETT is removed; the baby is again given Neopuff ventilation.*

BT-45 texts often lacked the kind of linking exemplified above. Hence, the relationship between intrinsic and extrinsic evaluation measures may need to account for both content and discourse structure.

Differences among users It is likely that some of the statistically weak results reported in the previous section are due to differences between participants in our evaluation. These differences are in part individual, as reflected by the comparatively high variance in performance scores (cf. Section 3). Perhaps more importantly, differences are also likely to arise between user groups. Although no main effect of Group was found on performance, it seems likely that different users will focus on different aspects of a text when reading summaries. For instance, whether a user is a doctor or a nurse will have an impact on how likely they are to consider taking a particular action, given their different aims, and the fact that some actions fall within their remit and others tend not to. Some support for this conclusion comes from the observation that the correlation between differences in performance and differences in informativeness and relevance were primarily due to one group of users, namely, Senior Doctors. If this interpretation is correct, then it reflects the necessity of tailoring the NLG system output to different users, particularly in a medical context, where roles tend to be fixed and a single 'one size fits all' solution is unlikely to be adequate [16].

Granularity Another limitation is related to the information captured by our annotations, which do not make sufficiently fine-grained distinctions and do not use the full expressivity available in the ontology. For example, *SaO2 decreases to 60%* and *SaO2 stays stable* are both represented by the same event, with TYPE = TREND and SOURCE = OXYGEN SATURATION; however, their relative importance is clearly different, since only the former suggests that something must be done. We are seeking to make finer-grained distinctions of this kind in our current work.

The role of context An issue which is related to granularity is the degree to which context must be taken into account. As shown in Table 2, human-authored texts displayed considerably higher variance in their informativeness and relevance scores compared to the BT-45 texts. BT-45's document planner had a relatively deterministic con-

tent selection strategy which selected events based on their clinical importance, computed using rules obtained from clinicians. By contrast, our expert authors may have been selecting content using much more sophisticated heuristics, in part based on the salient patterns in the data and possibly also their knowledge of the most important observations given a patient's current and previous state. As an example, BT-45 seldom mentioned noise or artifacts in the input signals, deeming these to be unimportant; in contrast, these were mentioned in some cases by the human texts because they drew a clinician's attention to the need for managing the sensors which were sampling the physiological parameters. In short, humans probably do a better job at taking context into account. Modulo differences in the probability of actions, our measures of informativeness and relevance gave equal weight to the different events mentioned, without reference to context, whereby an event can become relevant not through its association with a potential target action, but because it can shed light on the nature and provenance of the rest of the events described in the text. Thus, when we focus on those events which are indirectly linked to possible actions, BT-45 does not seem to differ very much from the H texts, but this should be interpreted in light of the fact that the two texts did differ on overall informativeness, as reflected by the NE measure.

8 Conclusions

This paper began by arguing that extrinsic evaluation methodologies, though useful and necessary, often leave open the question of which aspects of a system are contributing to the results, and why. The present paper attempted to identify some of these aspects, focusing primarily on the content selection strategy of a system to generate patient summaries in a Neonatal Intensive Care context. Our results showed that the relationship between our measures of textual content, and performance on a task is somewhat weak. Our interpretation of these results is that content-based intrinsic measures need to be more granular, and take into account other textual characteristics, such as discourse structure, as well as the role of different user groups. We have also proposed an intrinsic evaluation methodology which relies on domain knowledge to quantify informativeness and relevance. In our ongoing work, we are extending this methodology to address the shortcomings identified in our results.

Acknowledgments

Thanks to Professor Neil McIntosh, senior consultant at the Royal Infirmary of Edinburgh, who helped with the development of the ontology. Thanks also to Jim Hunter and Ehud Reiter. The BabyTalk project is supported by UK Engineering and Physical Sciences Research Council (EPSRC), under grants EP/D049520/1 and EP/D05057X/1.

References

[1] A. Belz and A. Gatt. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proc. ACL-08*, 2008.

[2] A. Belz and E. Reiter. Comparing automatic and human evaluation of NLG systems. In *Proc. EACL-06*, 2006.

[3] C. B. Callaway. Evaluating coverage for large symbolic NLG grammars. In *Proc. IJCAI-03*, 2003.

[4] C. B. Callaway and J. C. Lester. Narrative prose generation. *Artificial Intelligence*, 139(2):213–252, 2002.

[5] C. Calliston-Burch, M. Osborne, and P. Koehn. Re-evaluating the role of BLEU in machine translation research. In *Proc. EACL-06*, 2006.

[6] B. J. Dorr, C. Monz, S. President, R. Schwartz, and D. Zajic. A methodology for extrinsic evaluation of text summarization: Does ROUGE correlate? In *Proc. Workshop on Intrinsic and Extrinsic Evaluation Measures*, 2005.

[7] M. Foster. Automated metrics that agree with human judgements on generated output for an embodied conversational agent. In *Proc. INLG-08*, 2008.

[8] J. Hunter, G. Ewing, L. Ferguson, Y. Freer, R. Logie, P. McCue, and N. McIntosh. The NEONATE database. In *Proc. IDAMAP-03*, 2003.

[9] K. S. Jones and J. Galliers. *Evaluating natural language processing systems: An analysis and review*. Springer, Berlin, 1996.

[10] A. Karasimos and A. Isard. Multilingual evaluation of a natural language generation system. In *Proc. LREC-04*, 2004.

[11] I. Langkilde-Geary. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proc. INLG-02*, 2002.

[12] J. Lester and B. Porter. Developing and empirically evaluating robust explanation generators: The KNIGHT experiments. *Computational Linguistics*, 23(1):65–101, 1997.

[13] F. Portet, E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer, and C. Sykes. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7–8):789–816, 2009.

[14] E. Reiter, R. Robertson, and L. Osman. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144:41–58, 2003.

[15] O. Stock, M. Zancanaro, P. Busetta, C. Callaway, A. Krueger, M. Kruppa, T. Kuflik, E. Not, and C. Rocchi. Adaptive, intelligent presentation of information for the museum visitor in PEACH. *User Modeling and User-Adapted Interaction*, 17(3):257–304, 2007.

[16] B. Strople and P. Ottani. Can technology improve intershift report? what the research reveals. *Journal of Professional Nursing*, 22(3):197–204, 2006.

[17] M. van der Meulen, R. H. Logie, Y. Freer, C. Sykes, N. McIntosh, and J. Hunter. 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*, to appear.