Textual Properties and Task Based Evaluation: Investigating the Role of Surface Properties, Structure and Content.

Albert Gatt
Institute of Linguistics
University of Malta
albert.gatt@um.edu.mt

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

Abstract

This paper investigates the relationship between the results of an extrinsic, task-based evaluation of an NLG system and various metrics measuring both surface and deep semantic textual properties, including relevance. The latter rely heavily on domain knowledge. We show that they correlate systematically with some measures of performance. The core argument of this paper is that more domain knowledge-based metrics shed more light on the relationship between deep semantic properties of a text and task performance.

1 Introduction

Evaluation methodology in NLG has generated a lot of interest. Some recent work suggested that the relationship between various intrinsic and extrinsic evaluation methods (Spärck-Jones and Galliers, 1996) is not straightforward (Reiter and Belz, 2009; Gatt and Belz, to appear), leading to some arguments for more domain-specific intrinsic metrics (Foster, 2008). One reason why these issues are important is that reliable intrinsic evaluation metrics that correlate with performance in an extrinsic, task-based setting can inform system development. Indeed, this is often the stated purpose of evaluation metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin and Hovy, 2003), which were originally characterised as evaluation 'understudies'.

In this paper we take up these questions in the context of a knowledge-based NLG system, BT-45 (Portet et al., 2009), which summarises medical data for decision support purposes in a Neonatal Intensive Care Unit (NICU). Our extrinsic data comes from an experiment involving complex medical decision making based on automatically generated and human-authored texts (van der Meulen et al., 2009). This gives us the opportunity to directly compare the textual characteristics of generated and human-written summaries and their relationship to decision-making performance. The present work uses data from an earlier study (Gatt and Portet, 2009), which presented some preliminary results along these lines for the system in question. We extend this work in a number of ways. Our principal aim is to test the validity not only of general-purpose metrics which measure surface properties of text, but also of metrics which make use of domain knowledge, in the sense that they attempt to relate the 'deep semantics' of the texts to extrinsic factors, based on an ontology for the BT-45 domain.

After an overview of related work in section 2, the BT-45 system, its domain ontology and the extrinsic evaluation are described in section 3. The ontology plays an important role in the evaluation metrics presented in Section 5. Finally, the evaluation of the methods is presented in Section 6, before discussing and concluding in Section 7.

2 Related Work

In NLG evaluation, extrinsic, task-based methods play a significant role (Reiter et al., 2003; Karasimos and Isard, 2004; Stock et al., 2007). Depending on the study design, these studies often leave open the question of precisely which aspects of a system (and of the text it generates) contribute to success or failure. Intrinsic NLG evaluations often involve ratings of text quality or responses to questionnaires (Lester and Porter, 1997; Callaway and Lester, 2002; Foster, 2008), with some studies using post-editing by human experts (Reiter et al., 2005). Automatically computed metrics exploiting corpora, such as BLEU, NIST and ROUGE, have mainly been used in evaluations of the coverage and quality of morphosyntactic realisers (Langkilde-Geary, 2002; Callaway, 2003), though they have recently also been used

for subtasks such as Referring Expression Generation (Gatt and Belz, to appear) as well as end-to-end weather forecasting systems (Reiter and Belz, 2009). The widespread use of these metrics in NLP partly rests on the fact that they are quick and cheap, but there is controversy about their reliability both in MT (Calliston-Burch et al., 2006) and summarisation (Dorr et al., 2005; Liu and Liu, 2008). As noted in Section 1, similar questions have been raised in NLG. One of the problems associated with these metrics is that they rely on the notion of a 'gold standard', which is not always precisely definable given multiple solutions to the same generation, summarisation or translation task. These observations underlie recent developments in Summarisation evaluation such as the Pyramid method (Nenkova and Passonneau, 2004), which in addition also emphasises *content* overlap with a set of reference summaries, rather than n-gram matches.

It is interesting to note that, with some exceptions (Foster, 2008), most of the methodological studies on intrinsic evaluation cited here have focused on 'generic' metrics (corpus-based automatic measures being foremost among them), none of which use domain knowledge to quantify those aspects of a text related to its content. There is some work in Summarisation that suggests that incorporating more knowledge improves results. For example, Yoo and Song (Yoo et al., 2007) used the Medical Subject Headings (MeSH) to construct graphs representing the high-level content of documents, which are then used to cluster documents by topic, each cluster being used to produce a summary. In (Plaza et al., 2009), the authors have proposed a summarisation method based on WordNet concepts and showed that this higher level representation improves the summarisation task.

The principal aim of this paper is to develop metrics with which to compare texts using domain knowledge – in the form of the ontology used in the BT-45 system – and to correlate results to human decision-making performance. The resulting metrics focus on aspects of content, structure and relevance that are shown to correlate meaningfully with task performance, in contrast to other, more surface-oriented ones (such as ROUGE).

3 The BT-45 System

BT-45(Portet et al., 2009) was designed to generate a textual summary of 45 minutes of patient data in a Neonatal Intensive Care Unit (NICU), of the kind shown in Figure 1(a). The corresponding summary for the same data shown in Figure 1(b) is a two-step consensus summary written by two expert neonatologists. These two summaries correspond to two of the conditions in the task-based evaluation experiment described below.

In BT-45, summaries such as Figure 1(a) were generated from raw input data consisting of (a) physiological signals measured using sensors for various parameters (such as heart rate); and (b) discrete events logged by medical staff (e.g. drug administration). The system was based on a pipeline architecture which extends the standard NLG tasks such as document planning and microplanning with preliminary stages for data analysis and reasoning. The texts generated were descriptive, that is, they kept interpretation to a minimum (for example, the system did not make diagnoses). Nor were they generated with a bias towards specific problems or actions that could be considered desirable for a clinician to take in a particular context.

Every stage of the generation process made use of a domain-specific ontology of around 550 concepts, an excerpt of which is shown in Figure 1(c). The ontology classified objects of type EVENT and ENTITY into several subtypes; for example, a DRUG ADMINISTRATION is an INTERVENTION, which means it involves an agent and a patient. The ontology functioned as a repository of declarative knowledge, on the basis of which production rules were defined to support reasoning in order to make abstractions and to identify relations (such as causality) between events detected in the data based on their ontological class and their specific properties. In addition to the standard IS-A links, the ontology contains functional relationships which connect events to concepts representing physiological systems (such as the respiratory or cardiovascular systems); these are referred to as *functional concepts*. For example, in Figure 1(c), a FEED event is linked to NUTRITION, meaning that it is primarily relevant to the nutritional system. These links were included in the ontology following consultation with a senior neonatal consultant *after* the development of BT-45 was completed. Their inclusion was motivated by the knowledge-based evaluation metrics developed for the purposes of the present study, and discussed further

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.

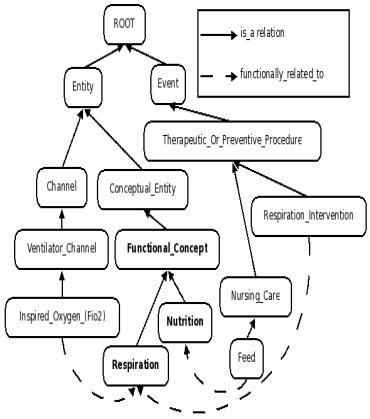
By 13:40 SaO2 had rapidly decreased to 79.

(a) BT-45 summary

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.

(b) Human summary



(c) Ontology excerpt

Figure 1: Excerpts from Human and BT-45 summaries, and ontology example.

in Section 5.

The task-based experiment to evaluate BT-45 was conducted off-ward and involved a group of 35 clinicians, who were exposed to 24 scenarios, each covering approximately 45 minutes of patient data, together with a short introductory text which gave some background about the patient. The patient data was then presented in one of three conditions: graphically using a time-series plot, and textually in the form of a consensus summary written by human experts (H; Figure 1(b)) and one generated automatically by BT-45(C; Figure 1(a)). Like the BT-45 texts, the H texts did not give interpretations or diagnoses and every effort was made not to bias a reader in favour of certain courses of action. A Latin Square design was used to ensure that each scenario was shown to an equal number of participants in each condition, while no participant saw the same scenario in more than one condition.

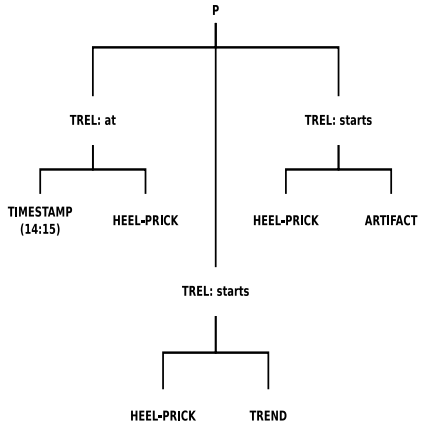
For each scenario, the task was to select one or more appropriate clinical actions from a predefined set of 18, one of which was 'no action'. Selections had to be made within three minutes, after which the scenario timed out. The same choice of 18 actions was given in each scenario s, but for each one, two neonatal experts identified the subsets of appropriate (AP_s), inappropriate/potentially harmful (INAP_s) and neutral actions. One of the appropriate actions was also deemed to be the 'target', that is, the most important action to take. In three scenarios, the 'target' was 'no action'. For each participant p and scenario s, the performance score P_s^p was based on the proportion P_{AP_s} of actions selected out of AP_s, and the proportion P_{INAP_s} selected out of the set of inappropriate actions INAP_s: $P_s^p = P_{\text{AP}_s} - P_{\text{INAP}_s} \in [-1, 1]$.

Overall, decision making in the H condition was better ($P_s = .45^{SD=.10}$) than either C ($P_s = .41^{SD=.13}$) or G ($P_s = .40^{SD=.15}$). No significant difference was found between the latter two, but the H texts were significantly better than the C texts, as revealed in a by-subjects ANOVA ($F(1, 31) = 5.266, p < 0.05$). We also performed a post-hoc analysis, comparing the proportions of appropriate actions selected, P_{AP} and that of inappropriate actions P_{INAP} in the H and C conditions across scenarios. In addition, we computed a different score SP_{AP}, defined as the proportion of appropriate actions selected by a participant within a scenario out of the total number of actions selected (effectively a measure of 'precision'). A comparison between means for these three scores obtained across scenarios showed no significant differences.

In the analysis reported in Section 6, we compare our textual metrics to both the global score P as well as to these three other performance indicators. In various follow-up analyses (van der Meulen et al., 2009; Reiter et al., 2008), it was found that the three scenarios in which the target action was 'no action' may have misled some participants, insofar as this option was included among a set of other actions, some of which were themselves deemed appropriate or at least neutral (in the sense that they could be carried out without harming the patient). We shall therefore exclude these scenarios from our analyses.

```
<P>
 At 14:15 hours

 <EVENT TYPE="HEEL_PRICK" ID="e11">
  a heel prick is done.
 </EVENT>

 <EVENT TYPE="TREND" SOURCE="HR" DIRECTION="increasing" ID="e12">
  The HR increases
 </EVENT>

 at this point and for 7 minutes from the start of this procedure

 <EVENT CARDINALITY="3" SOURCE="SaO2" TYPE="ARTIFACT" ID="e13">
  there is a lot of artefact in the oxygen saturation trace.
 </EVENT>
</P>

<TREL ARG0="e11" ARG1="TIMESTAMP" RELATION="at" />
<TREL ARG0="e12" ARG1="e11" RELATION="starts" />
<TREL ARG0="e13" ARG1="e11" RELATION="starts" />
```

(a) Annotation



(b) Normalised tree

Figure 2: Fragment of an annotated summary and normalised tree representation.

4 Corpus Annotation

For this study, we annotated the H and C texts from our experiment using the ontology, in order to make both their semantic content and structure explicit. Figure 2(a) shows an excerpt from an annotated text. Every paragraph of the text is marked up explicitly. All segments of the text corresponding to an ontology EVENT are marked up with a TYPE (the name of the concept in the ontology) and other properties, such as DIRECTION and SOURCE in the case of trends in physiological parameters. The CARDINALITY attribute is used to indicate that a single text segment abstracts over several occurrences in the data; for example, the statement about artefacts in the example corresponds to three such episodes in the data.

In addition to events, the markup also includes separate nodes for all the temporal (TREL) and discourse (DREL) relations which are explicitly mentioned in the text, typically using adverbial or prepositional phrases or verbs of causality. Every TREL and DREL points to two arguments and has a RELATION attribute. In the case of a TREL, the value is one of the temporal relations defined by (Allen, 1983). For DRELs, values were restricted to CAUSE and CONTRAST (Mann and S.Thompson, 1988). One of the arguments of a TREL can be a `timestamp`, rather than an event. This is the case for the first sentence in the fragment, where event `e11` is specified as having occurred at a specific time (*at 14:15*). By contrast, `r4` is a relation between `e11` and `e12`, where the RELATION is STARTS, indicating that the text specifies that `e11` is used by the author as the anchor point to specify the start of `e12`, as reflected by the expression *at this point*.

The markup provided the basis on which many of the metrics described in the following section were computed. Based on the annotation, we used a normalised structural representation of the texts as shown in Figure 2(b), consisting of PARAGRAPH (P) nodes which subsume events and relations. Relations dominate their event arguments. For example, the `starts` TREL holding between `e12` and `e11` is represented by a STARTS node subsuming the two events. In case an event is dominated by more than one relation (for example, it is temporally related to two events, as `e11` is in Figure 2(a), we maintain the tree structure by creating two copies of the event, which are subsumed by the two relations. Thus, the normalised tree representation is a 'compiled out' version of the graph representing all events and their relations. The tree representation is better suited to our needs, given the complexity of comparing two graphs.

5 Metrics

The evaluation metrics used to score texts written by domain experts and those generated by the BT-45 system fall into three main classes, described below.

Semantic content and structure To compare both the content and the structure of texts, we used three measures. The first quantifies the number of EVENT nodes in an annotated text, defined as

$\sum_{e \in \text{E}} c$, where E is the set of events mentioned, and c is the value of the CARDINALITY attribute of an event $e \in$ E. Similarly, we computed the number of temporal (TREL) and discourse (DREL) relations mentioned in a text. We also used the Tree Edit Distance metric to compute the distance between the tree representations of the H and C texts (see Figure 2(b)). This measure computes the minimum number of node insertions, deletions and substitutions required to transform one tree into another and therefore takes into account not only the content (events and relations) but also its structural arrangement in the text. The edit distance between two trees is computed using the standard Levenshtein edit distance algorithm, computed over a string that represents the preorder traversal of the two trees, using a cost of 1 for insertions and deletions, and 2 for substitutions.

N-gram overlap As a measure of n-gram overlap, we use ROUGE-n, which measures simple $n-$gram overlap (in the present paper we use $n = 4$). We also use ROUGE-SU, in which overlap is computed using skip-bigrams while also accounting for unigrams that a text has in common with its reference (in order to avoid bias against texts which share several unigrams but few skip bigrams).

Domain-dependent relevance metrics The metrics described so far make use of domain knowledge only to the extent that this is reflected in the textual markup. We now consider a family of metrics which are much more heavily reliant on domain-specific knowledge structures and reasoning. In our domain, the relevance of a text in a given experimental scenario s can be defined in terms of whether the events it mentions have some relationship to the appropriate clinical actions (AP_s). We attempt to model some aspects of this using a weighting strategy and reasoning rules.

Recall from Section 3 that fc's represent the various physiological systems to which an event or action can be related. Therefore, each event e mentioned in a text can be related to a set of possible actions using the functional concepts $fc(e)$ to which that event is linked in the ontology. Let $\text{E}_{s,t}$ be the set of events mentioned in text t for scenario s. An event $e \in \text{E}_{s,t}$ *references* an action a iff $\text{FC}(e) \cap \text{FC}(a) \neq \emptyset$. Our hypothesis is that an appropriate action is more likely to be taken if there are events which reference it in the text – that is, if the text mentions things which are directly or indirectly *relevant* to the action. For instance, if a text mentions events related to the RESPIRATION fc, a clinician might be more likely to make a decision to manage a patient's ventilation support. It is worth emphasising that, since both the BT-45 and human-authored texts were descriptive and were not written or generated with the appropriate actions in mind, the hypothesis that the relevance of the content to the appropriate actions might increase the likelihood of these actions being chosen is far from a foregone conclusion.

Part of the novelty in this way of quantifying relevance lies in its use of the knowledge (i.e. the ontology) that is already available to the system, rather than asking human experts to rate the relevance of a text, a time-consuming process which could be subject to experts' personal biases. However, this way of conceptualising relevance generates links to too many actions for one event. It is often the case that an event, through its association with a functional concept, references more than one action, but not all of these are appropriate. For example, a change in oxygen saturation can be related to RESPIRATION, which itself is related to several respiration-related actions in a scenario, only some of which are appropriate. Clearly, relevance depends not only on a physiological connection between an event and a phsiological system (functional concept), but also on the *context*, that is, the other events and their relative importance in a given scenario. Another factor that needs to be taken into account is the overall probability of an action. Some actions are performed routinely, while others tend to be associated with emergencies (e.g. a nappy change is much more frequent over all than resuscitating a patient). This means that some actions – even appropriate ones – may have been less likely to be selected even though they were referenced by the text and were appropriate.

We prune unwarranted connections between events and actions by taking into account (a) a patient's current status (described in the text and in the background information given to experimental participants); (b) the fact that some actions have much higher prior probabilities than others because they are performed more routinely; (c) the fact that some events may be more important than others (e.g. resuscitation is much more important

than a nappy change). Based on this, we define the *weight* of an action a as follows:

$$W_a = \frac{\sum_{e \in \mathrm{E}} \frac{Pr(a) * e.importance}{\sum_{a \in \mathrm{A}_e} Pr(a)}}{\sum_{e \in \mathrm{E}} e.importance} \quad (1)$$

Where E is the set of events in the text, A_e the set of actions related to event e, $e.importance \in \mathbb{N}^+$ the importance of the event e and $Pr(a)$ the prior probability of action a. All weights are normalised so that the following inequalities hold:

$$\sum_{a \in \mathrm{A}_e} \frac{Pr(a) * e.importance}{\sum_{a \in \mathrm{A}_e} Pr(a)} = e.importance \quad (2)$$

$$\sum_{a \in \mathrm{A}} W_a = 1 \quad (3)$$

where A is the set of all possible actions. The idea is that an event e makes some contribution (possibly 0) to the relevance of some actions A_e, and the total weight of the event is distributed among all actions related to it using (a) the prior probability $Pr(a)$ of each action (the most frequent action will have more weight) and (b) the importance of the event. At the end of the process each action would be assigned a score representing the accumulated weights of the events, which is then normalised, so that $\sum_{a \in \mathrm{A}} W_a = 1$.

The prior probability in the equation is meant to reflect our earlier observation that clinical actions differ in the frequency with which they are performed and this may bias their selection. Priors were computed using maximum likelihood estimates from a large database containing exhaustive annotations of clinical actions recorded by an on-site research nurse over a period of 4 months in a NICU, which contains a total of 43,889 records of actions (Hunter et al., 2003).

The importance value in equation (1) is meant to reflect the fact that events in the text do not attract the attention of a reader to the same extent, since they do not have the same degree of 'severity' or 'surprise'. We operationalise this by identifying the superconcepts in the ontology (PATHOLOGICAL-FUNCTION, DISORDER, SURGICAL-INTERVENTION, etc.) which could be thought of as representing 'drastic' occurrences. To these we added the concept of a TREND which corresponds to a change in a physiological parameter (such as an increase in heart rate), based on the rationale that the primary aim of NICU staff is to keep a patient stable, so that any physiological instability warrants an intervention. The importance of events subsumed by these superconcepts was then set to be three times that of 'normal' events.

Finally, we apply knowledge-based rules to prune the number of actions A_e related to an event e. As an example, a decision to intubate a baby depends not only on events in the text which reference this action, but also on whether the baby is already intubated. This can be assessed by checking whether s/he is on CMV (a type of ventilation which is only used after intubation). The rule is represented as $\text{INTUBATE} \rightarrow \neg\text{on}(\text{baby}, \text{CMV})$. Although such rules are extremely rough, they do help to prune inconsistencies.

Two scores were computed for both human and computer texts using equation (1). $\text{REL}_{s,t}$ is the sum of weights of actions referenced in a text t for scenario s which are appropriate: $\text{REL}_{s,t} = \sum_{a \in \mathrm{A}_{ap}} W_a$. Conversely, $\text{IRREL}_{s,t}$ quantifies the weights of actions referenced in t for scenario s which are inappropriate: $\text{IRREL}_{s,t} = \sum_{a \in \mathrm{A}_{inap}} W_a$.

6 Results

In what follows, we report two-tailed Pearson's r correlations to compare our metrics to the three performance measures discussed in Section 3: P, the global performance score; P_{APP} and P_{INAPP}, the proportion of appropriate (resp. inappropriate) actions selected from the subsets of in/appropriate (resp. inappropriate) actions in a scenario; and SP_{APP}, the proportion of appropriate actions selected by a participant out of the set of actions selected. The last three are included because they shed light more directly on the extent to which experimental participants chose correctly or incorrectly. In case a metric measures similarity or difference between texts, the correlation reported is with the *difference* between the H scores and the C scores. Where relevant, we also report correlations with the *absolute* mean performance scores within the H and/or C conditions. Correlations exclude the three scenarios which had 'no action' as the target appropriate action, though where relevant, we will indicate whether the correlations change when these scenarios are also included.

6.1 Content and Structure

Overall, the C texts mentioned significantly fewer events than the H texts ($t_{20} = 2.44, p = .05$), and also mentioned fewer temporal and discourse relations explicitly ($t_{20} = 3.70, p < .05$). In

	P (H-C)	P_{APP} (H-C)	SP_{APP} (H-C)	P_{INAP} (H-C)
Events (H-C)	.43$^{\diamondsuit}$	.42$^{\clubsuit}$	.02	-.09
Relations (H-C)	.34	.30	0	-.15
Tree Edit	.36	.33	.09	-.14

Table 1: Correlations between performance differences and content/structure measures. $^{\diamondsuit}$significant at $p = .05$; $^{\clubsuit}$approaches significance at $p = .06$

the case of the H texts, the number of events and relations did not correlate significantly with any of the performance scores. In the case of the C texts, the number of relations mentioned was significantly negatively correlated to P_{INAP} ($r = -.49, p < .05$), and positively correlated to SP_{APP} ($r = .7, p < .001$). This suggests that temporal and discourse relations made texts more understandable and resulted in more appropriate actions being taken. More unexpectedly, the number of events mentioned was negatively correlated to P_{APP} ($r = -.53, p < .05$) and to P ($r = -.5, p < .05$). This may have been due to the C texts mentioning a number of events that were relatively unimportant and/or irrelevant to the appropriate actions.

Table 1 displays correlations between performance differences between H and C, and differences in number of events and relations, as well as Tree Edit Distance. The positive correlation between the number of events mentioned and P suggests that a larger amount of content in the H texts is partially responsible for the difference in decision-making accuracy by experimental participants. This is further supported by the fact that the correlation with the difference in P_{APP} approaches significance. It is worth noting that none of these correlations are significant when means from the three 'no action' scenarios are included in the computation. This further supports our earlier conclusion that these three scenarios are outliers. Somewhat surprisingly, Tree Edit Distance does not correlate significantly with any of the performance differences, though the correlations go in the expected directions (positive in the case of P, SP_{APP} and P_{APP}, negative in the case of P_{INAP}). This may be due to the high variance in the Edit Distance scores (mean: 66.5; SD: 34.8).

Overall, these results show that differences in both content and structure made the H texts superior and human texts did a much better job at explicitly relating events or situating them in time, which is crucial for comprehension and correct decision-making. This point has previously been made in relation to the same data on the basis of a qualitative study (Reiter et al., 2008).

	Absolute Scores (C)				**Differences (H-C)**			
	P	P_{AP}	P_{INAP}	SP_{AP}	P	P_{AP}	P_{INAP}	SP_{AP}
R-4	.33	.38	.2	-.03	-.19	-.2	-.01	-.1
R-SU	-.03	-.02	.05	-.31	.04	.01	-.1	.13

Table 2: Correlations between ROUGE and performance scores in the C condition. $^{\diamondsuit}$significant at $p = .05$.

6.2 N-gram Overlap

Correlations with ROUGE-4 and ROUGE-SU are shown in Table 2 both for absolute performance scores on the C texts, and for the differences between H and C. This is because ROUGE can be interpreted in two ways: on the one hand, it measures the 'quality' of C texts relative to the reference human texts; on the other it also indicates similarity between C and H.

There are no significant correlations between ROUGE and any of our performance measures. Although this leaves open the question of whether a different set of performance measures, or a different experiment, would evince a more systematic covariation, the results suggest that it is not surface similarity (to the extent that this is measured by ROUGE) that is contributing to better decision making. It is however worth noting that some correlations with ROUGE-4, namely those involving P and P_{APP}, do turn out significant when the 'no action' scenarios are included. This turns out to be solely due to one of the 'no action' scenarios, which had a much higher ROUGE-4 score than the others, possibly because the corresponding human text was comparatively brief and the number of events mentioned in the two texts was roughly equal (11 for the C text, 12 for the H text).

6.3 Knowledge Based Relevance Metrics

Finally, we compare our knowledge-based measures of the relevance of the content to appropriate actions (REL) and to inappropriate actions (IRREL). The correlations between each measure and

	Human (H)				BT-45 (C)			
	P	P_{AP}	P_{INAP}	SP_{AP}	P	P_{AP}	P_{INAP}	SP_{AP}
REL	.14	.11	-.14	.60$^{\diamond}$	.33	.24	-.49$^{\diamond}$	.7$^{\diamond}$
IRREL	-.25	-.22	.1	-.56$^{\diamond}$	-.34	-.26	.43	-.62$^{\diamond}$

Table 3: Correlations between knowledge-based relevance scores and absolute performance scores in the C and H conditions. $^{\diamond}$significant at $p \leq .05$.

the absolute performance scores in each condition are displayed in Table 3.

The absolute scores in Table 3 show that both REL and IRREL are significantly correlated to SP_{APP}, the proportion of appropriate actions out of the actions selected by participants. The correlations are in the expected direction: there is a strong tendency for participants to choose more appropriate actions when REL is high, and the reverse is true for IRREL. In the case of the C texts, there is also a negative correlation (as expected) between REL and P_{INAP}, though this is the only one that reaches significance with this variable. It therefore appears that the knowledge-based relevance measures evince a meaningful relationship with at least some of the more 'direct' measures of performance (those assessing the relative preference of participants for appropriate actions based on a textual summary), though not with the global preference score P. One possible reason for the low correlations with the latter is that the two measures attempt to quantify directly the relevance of the content units in a text to in/appropriate courses of action; hence, they have a more direct relationship to measures of proportions of the courses of actions chosen.

7 Discussion and Conclusions

We conclude this paper with some observations about the relative merit of different measures of textual characteristics. 'Standard', surface-based measures such as (ROUGE) do not display any systematic relationship with our extrinsic measures of performance, recalling similar observations in the NLG literature (Gatt and Belz, to appear) and in MT and Summarisation (Calliston-Burch et al., 2006; Dorr et al., 2005). Some authors have also reported that ROUGE does not correlate well with human judgements of NLG texts (Reiter and Belz, 2009). On the other hand, we do find some evidence that the amount of content in texts, and the extent to which they explicitly relate content elements temporally and rhetorically, may impact decision-making. The significant correlations observed between the number of relations in a text and the extrinsic measures are worth emphasising, as they suggest a significant role not only for content, but also rhetorical and temporal structure, something that many metrics do not take into account.

Perhaps the most important contribution of this paper has been to emphasise knowledge-based aspects of textual evaluation, not only by measuring content units and structure, but also by developing a motivated *relevance* metric, the crucial assumption being that the utility of a summary is contingent on its managing to convey information that will motivate a reader to take the 'right' course of action. The strong correlations between the relevance measures and the extent to which people chose the correct actions (or more accurately, chose *more* correct actions) vindicates this assumption.

Some of the correlations which turned out not to be significant may be due to 'noise' in the data, in particular, high variance in the performance scores (as suggested by the standard deviations for P given in Section 3). They therefore do not warrant the conclusion that *no* relationship exists between a particular measure and extrinsic task performance; nevertheless, where other studies have noted similar gaps, the trends in question may be systematic and general. This, however, can only be ascertained in further follow-up studies.

This paper has investigated the relationship between a number of intrinsic measures of text quality and decision-making performance based on an external task. Emphasis was placed on metrics that quantify aspects of semantics, relevance and structure. We have also compared generated texts to their human-authored counterparts to identify differences which can motivate further system improvements. Future work will focus on further exploring metrics that reflect the relevance of a text, as well as the role of temporal and discourse structure in conveying the intended meaning.

References

J. F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Charles B. Callaway and James C. Lester. 2002. Narrative prose generation. *Artificial Intelligence*, 139(2):213–252.

C. B. Callaway. 2003. Evaluating coverage for large symbolic NLG grammars. In *Proc. IJCAI'03*.

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

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

M.E. Foster. 2008. Automated metrics that agree with human judgements on generated output for an embodied conversational agent. In *Proc. INLG'08*.

A. Gatt and A. Belz. to appear. Introducing shared task evaluation to NLG: The TUNA shared task evaluation challenges. In E. Krahmer and M. Theune, editors, *Empirical Methods in Natural Language Generation*. Springer.

A. Gatt and F. Portet. 2009. Text content and task performance in the evaluation of a natural language generation system. In *Proc. RANLP'09*.

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

A. Karasimos and A. Isard. 2004. Multilingual evaluation of a natural language generation system. In *Proc. LREC'04*.

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

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

C-Y Lin and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. in. In *Proc. of HLT-NAACL'03*.

F. Liu and Y. Liu. 2008. Correlation between rouge and human evaluation of extractive meeting summaries. In *Proc. ACL'08*.

W. C. Mann and S.Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organisation. *Text*, 8(3):243–281.

A. Nenkova and R. Passonneau. 2004. Evaluating content selection in summarisation: The Pyramid method. In *Proc. NAACL-HLT'04*.

S. Papineni, T. Roukos, W. Ward, and W. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proc. ACL'02*.

L Plaza, A Díaz, and P Gervás P. 2009. Automatic summarization of news using wordnet concept graphs. best paper award. In *Proc. IADIS'09*.

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

E. Reiter and A. Belz. 2009. An investigation into the validity of some metrics for automatically evaluating Natural Language Generation systems. *Computational Linguistics*, 35(4):529–558.

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

E. Reiter, S. Sripada, J. Hunter, J. Yu, and I. Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167:137–169.

E. Reiter, A. Gatt, F. Portet, and M. van der Meulen. 2008. The importance of narrative and other lessons from an evaluation of an NLG system that summarises clinical data. In *Proc. INLG'08*.

K. Spärck-Jones and J. R. Galliers. 1996. *Evaluating natural language processing systems: An analysis and review*. Springer, Berlin.

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

M. van der Meulen, R. H. Logie, Y. Freer, C. Sykes, N. McIntosh, and J. Hunter. 2009. When a graph is poorer than 100 words. *Applied Cognitive Psychology*, 24(1):77–89.

I. Yoo, X. Hu, and I-Y Song. 2007. A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. *BMC Bioinformatics*, 8(9).