Attribute Selection for Referring Expression Generation: New Algorithms and Evaluation Methods

Albert Gatt
Department of Computing Science
University of Aberdeen
Aberdeen AB24 3UE, UK
a.gatt@abdn.ac.uk

Anja Belz
Natural Language Technology Group
University of Brighton
Brighton BN2 4GJ, UK
a.s.belz@brighton.ac.uk

Abstract

Referring expression generation has recently been the subject of the first Shared Task Challenge in NLG. In this paper, we analyse the systems that participated in the Challenge in terms of their algorithmic properties, comparing new techniques to classic ones, based on results from a new human task-performance experiment and from the intrinsic measures that were used in the Challenge. We also consider the relationship between different evaluation methods, showing that extrinsic task-performance experiments and intrinsic evaluation methods yield results that are not significantly correlated. We argue that this highlights the importance of including extrinsic evaluation methods in comparative NLG evaluations.

1 Introduction and Background

The Generation of Referring Expressions (GRE) is one of the most intensively studied sub-tasks of Natural Language Generation (NLG). Much research has focused on content determination in GRE, which is typically framed as an *attribute selection* task in which properties are selected in order to describe an intended referent. Since such properties are typically defined as attribute-value pairs, in what follows, we will refer to this as the ASGRE (*A*ttribute *S*election for *G*enerating *R*eferring *E*xpressions) task.

1.1 Approaches to ASGRE

Since early work on ASGRE, which focused on pragmatic motivations behind different types of reference (Appelt, 1985; Appelt and Kronfeld, 1987), the focus has increasingly been on definite descriptions and identification, where the set of attributes selected should uniquely distinguish the intended referent from other entities (its 'distractors'). *Unique Reference* in this sense is a dominant criterion for selecting attribute sets in classic ASGRE algorithms. Following Dale (1989), and especially Dale and Reiter (1995), several contributions have extended the remit of ASGRE algorithms to handle relations (Dale and Haddock, 1991; Kelleher and Kruijff, 2006) and gradable attributes (van Deemter, 2006); and also to guarantee logical completeness of algorithms (van Deemter, 2002; Gardent, 2002; Horacek, 2004; Gatt and van Deemter, 2007).

Much of this work has incorporated the principle of brevity. Based on the Gricean Quantity maxim (Grice, 1975), and originally discussed by Appelt (1985), and further by Dale (1989), Reiter (1990) and Gardent (2002), this principle holds that descriptions should contain no more information than is necessary to distinguish an intended referent. In ASGRE, this has been translated into a criterion which determines the adequacy of an attribute set, implemented in its most straightforward form in *Full Brevity* algorithms which select the smallest attribute set that uniquely refers to the intended referent (Dale, 1989).

Another frequent property of ASGRE algorithms is *Incrementality* which involves selection of attributes one at a time (rather than exhaustive search for a distinguishing set), and was initially motivated by algorithmic complexity considerations. The Incremental Algorithm of Dale and Reiter (1995) was also justified with reference to psycholinguistic findings that (a) humans overspecify their references (i.e. they are not brief when they produce a refer-

ence); and (b) this tends to occur with properties which seem to be very salient or in some sense preferred (Pechmann, 1989; Belke and Meyer, 2002; Arts, 2004). In the Incremental Algorithm (Dale and Reiter, 1995), incrementality took the form of hillclimbing along an attribute order which reflects the preference that humans manifest for certain attributes (such as COLOUR). This *Modelling of Human Preferences* in the Incremental Algorithm has proven influential. Another feature proposed by Dale and Reiter is to hardwire the inclusion of the type of an entity in a description, reflecting the human tendency to always include the category of an object (e.g. *chair* or *man*), even if it has no discriminatory value. A compromise may be reached between incrementality and brevity; for example, Dale's (1989) Greedy Algorithm selects attributes one at a time, and computes the *Discriminatory Power* of attributes, that is, it bases selection on the extent to which an attribute helps distinguish an entity from its distractors.

In the remainder of this paper, we uniformly use the term 'algorithmic property' for the selection criteria and other properties of ASGRE algorithms described above, and refer to them by the following short forms: Full Brevity, Uniqueness, Discriminatory Power, Hardwired Type Selection, Human Preference Modelling and Incrementality.[1]

1.2 ASGRE and Evaluation

Though ASGRE evaluations have been carried out (Gupta and Stent, 2005; Viethen and Dale, 2006; Gatt et al., 2007), these have focused on 'classic' algorithms, and have been corpus-based. The absence of task-performance evaluations is surprising, considering the well-defined nature of the ASGRE task, and the predominance of task-performance studies elsewhere in the NLG evaluation literature (Reiter et al., 2003; Karasimos and Isard, 2004).

Given the widespread agreement on task definition and input/output specifications, ASGRE was an ideal candidate for the first NLG shared task evaluation challenge. The challenge was first discussed during a workshop held at Arlington, Va. (Dale and White, 2007), and eventually organised as part of the UCNLG+MT Workshop in September 2007 (Belz and Gatt, 2007).

The ASGRE Shared Task provided an opportunity to (a) assess the extent to which the field has diversified since its inception; (b) carry out a comparative evaluation involving both automatic methods and human task-performance methods.

1.3 Overview

In the ASGRE Challenge report (Belz and Gatt, 2007) we presented the results of the ASGRE Challenge evaluations objectively and with little interpretation. In this paper, we present the results of a new task-performance evaluation and a new intrinsic measure involving the same 15 systems and compare the new results with the earlier ones. We also examine all results in the light of the algorithmic properties of the participating systems. We thus focus on two issues in ASGRE and its evaluation. We examine the similarities and differences among the systems submitted to the ASGRE Challenge and compare them to classic approaches (Section 2.1). We look at the results of intrinsic and extrinsic evaluations (Section 4) and examine how these relate to algorithmic properties (Section 4.1 and 4.2). Finally we look at how intrinsic and extrinsic evaluations correlate with each other (Section 4.3).

2 The ASGRE Challenge

The ASGRE Challenge used the TUNA Corpus (Gatt et al., 2007), a set of human-produced referring expressions (REs) for entities in visual domains of pictures of furniture or people. The corpus was collected during an online elicitation experiment in which subjects typed descriptions of a target referent in a DOMAIN in which there were also 6 other entities ('distractors'). Each RE in the corpus is paired with a domain representation consisting of the target referent and the set of distractors, each with their possible attributes and values; Figure 1(a) shows an example of an entity representation.

In each human-produced RE, substrings are annotated with the attribute(s) they express ('realise'). For the Challenge, the attributes in each RE were extracted to produce a DESCRIPTION, i.e. a set of

[1] While terms like 'Unique Reference' and 'Brevity' are characteristics of outputs of ASGRE algorithms, we use them as shorthand here for those properties of algorithms which guarantee that they will achieve these characteristics in the attribute sets generated.

```
<ENTITY ID="121" TYPE="target">
<ATTRIBUTE NAME="colour" VALUE="blue" />
<ATTRIBUTE NAME="orientation" VALUE="left" />
<ATTRIBUTE NAME="type" VALUE="fan" />
<ATTRIBUTE NAME="size" VALUE="small" />
<ATTRIBUTE NAME="x-dimension" VALUE="1" />
<ATTRIBUTE NAME="y-dimension" VALUE="3" />
</ENTITY>
```

(a) Entity representation

```
<DESCRIPTION>
<ATTRIBUTE NAME="colour" VALUE="blue" />
<ATTRIBUTE NAME="orientation" VALUE="left" />
<ATTRIBUTE NAME="type" VALUE="fan" />
</DESCRIPTION>
```

(b) Description: *the blue fan facing left*

Figure 1: Example of the input and output data in the ASGRE Challenge

attribute-value pairs to describe the target referent. An example, corresponding to a human-produced RE for Figure 1(a), is shown in Figure 1(b).

For the ASGRE Challenge, the 780 singular descriptions in the corpus were used, and divided into 60% training data, 20% development data and 20% test data. Participants were given both input (DOMAIN) and output (DESCRIPTION) parts in the training and development data, but just inputs in the test data. They were asked to submit the corresponding outputs for test data inputs.

2.1 Systems

The evaluations reported below included 15 of the 22 submitted ASGRE systems. Table 1 is an overview of these systems in terms of five of the classic algorithmic properties described in Section 1, and one property that has emerged in more recent ASGRE algorithms: *Trainability*, or automatic adaptability of systems from data. This classification of systems is based on the reports submitted by their developers to the ASGRE Challenge. Properties are indicated in the first column (abbreviations are explained in the table caption).

The version of the IS-FBS system that was originally submitted to ASGRE contained a bug and did not actually output minimal attribute sets (but added an arbitrary attribute to each set). Unlike the ASGRE Challenge task-performance evaluation, the analysis presented in this paper uses the corrected version of this system.

3 Evaluation methods

Evaluation methods can be characterised as either *intrinsic* or *extrinsic*. While intrinsic methods evaluate the outputs of algorithms in their own right, either relative to a corpus or based on absolute evaluation metrics, extrinsic methods assess the effect of an algorithm on something external to it, such as its effect on human performance on some external task.

3.1 Intrinsic measures

In the evaluation results reported below, we use two of the intrinsic methods used in the ASGRE Challenge. The first of these is *Minimality*, defined as the proportion of descriptions produced by a system that are maximally brief, as per the original definition in Dale (1989). The second is the Dice coefficient, used to compare the description produced by a system (D_S) to the human-produced description (D_H) on the same input domain. Dice is estimated as follows:

$$\frac{2 \times |D_S \cap D_H|}{|D_S| + |D_H|} \quad (1)$$

For this paper, we also computed MASI, a version of the Jaccard similarity coefficient proposed by Passonneau (2006) which multiplies the similarity value by a monotonicity coefficient, biasing the measure towards those cases where D_S and D_H have an empty set difference. Intuitively, this means that those system-produced descriptions are preferred which do not include attributes that are omitted by a human. Thus, two of our intrinsic measures assess *Humanlikeness* (Dice and MASI), while *Minimality* reflects the extent to which an algorithm conforms to brevity, one of the principles that has emerged from the ASGRE literature.

3.2 Extrinsic measures

In the ASGRE Challenge, the extrinsic evaluation was performed via an experiment in which partici-

	CAM					GRAPH		IS				TITCH			
	T	TU	B	BU	DIT-DS	FP	SC	FBS	FBN	IAC	NIL	RS	RS+	AS	AS+
Incr	Y	Y	Y	Y	Y	n	n	n	n	Y	Y	Y	Y	Y	Y
DP	Y	Y	Y	Y	n	n	n	n	n	n	n	Y	Y	Y	Y
Train	Y	Y	n	n	Y	Y	Y	n	Y	Y	Y/n	Y	Y	Y	Y
Type	Y	Y	Y	Y	Y	Y	Y	n	n	n	n	Y	Y	Y	Y
Hum	Y	Y	Y	Y	n	Y	n	n	n	n	n	n	n	n	n
FB	n	n	n	n	n	n	n	Y	n	n	n	n	n	n	n

Table 1: Overview of properties of systems submitted to the ASGRE Challenge. All systems produce outputs with unique reference. The meaning of the abbreviations is as follows (for definitions see Section 1): *Incr = Incrementality*; *DP = Discriminatory Power*; *Train = Trainability*; *Type = Hardwired Type Selection*; *Hum = Human Preference Modelling*; *FB = Full Brevity*

.

pants were shown visual representations of the domains that were used as inputs in the test data, coupled with a system-generated description on the same screen. Their task was to identify the intended referent in the domain by means of a mouse click. The rationale behind the experiment was to assess the degree to which an algorithm achieved its stated purpose of generating a description that facilitates identification of the intended referent.

Since system outputs were sets of attributes (see Figure 1(b)), they were first mapped to NL strings using a deterministic, template-based method which always maps each attribute to word(s) in the same way and in the same position regardless of context[2].

The present analysis is based on results from a new evaluation experiment which replicated the original methodology with a slight difference. While participants in the ASGRE experiment were shown visual domains and descriptions at the same time, this experiment sought to distinguish reading and identification time. Thus, for each system output, participants first saw the description, using the mouse to call up the visual domain once they had read it. This yields three dependent measures: (a) reading time (RT); (b) identification time (IT); (c) error rate (ER), defined as the proportion of trials per system for which participants identified the wrong referent.

One of the possible complications with the RT measure is that, while it depends on the semantics of a description (the attributes that an algorithm selects), the syntactic complexity of the description will also impact the time it takes to read it. In our setup, the adoption of a deterministic, one-to-one mapping between an attribute and its linguistic representation controls for this to some extent. Since every attribute in any semantic representation will invariably map to the same surface form, there is no variation between systems in terms of how a particular attribute is realised.

Design: As in the original experiment, we used a Repeated Latin Squares design in which each combination of peer system and test set item is allocated one trial. Because there were 148 items in the test set, but 15 peer systems, 2 test set items were randomly selected and duplicated to give a test set size of 150, and 10 Latin Squares. The 150 test set items were divided into two sets of 75; 15 of the 30 participants did the first 75 items (the first 5 Latin Squares), while the other 15 did the rest. This resulted in 2250 trials (each corresponding to a combination of test set item and system) in all. For the purposes of this paper, the duplicate items were treated as fillers, that is, every item was included only once in the analysis.

Participants and procedure: The experiment was carried out by 30 participants recruited from among the faculty and administrative staff of the University of Brighton. Participants carried out the experiment under supervision in a quiet room on a laptop. Stimulus presentation was carried out using DMDX, a Win-32 software package for psycholinguistic experiments involving time measurements (Forster and Forster, 2003). Participants initiated each trial, which consisted of an initial warning bell and a fixation point flashed on the screen for 1000ms. They then read the description and called up the visual domain to identify the referent. Identi-

[2]The template-based string-mapping algorithm was created by Irene Langkilde-Geary at the University of Brighton.

fication was carried out by clicking on the image that a participant thought was the target referent of the description they had read. RT was measured from the point at which the description was presented, to the point at which a participant called up the next screen via mouse click. IT was measured from the point at which pictures (the visual domain) were presented on the screen to the point where a participant identified a referent by clicking on it. In addition to the time taken to identify a referent, the program recorded the location of a participant's mouse click, in order to assess whether the object identified was in fact the correct one. Trials timed out after 15 seconds; this only occurred in 2 trials overall (out of 2250).

4 Results and discussion

Table 2 displays aggregate (mean or percentage) scores for each of the intrinsic and extrinsic measures. As an initial test for differences between the 15 systems separate univariate ANOVAs were conducted using SYSTEM as the independent variable (15 levels), testing its effect on the extrinsic task-performance measures (RT and IT). For error rate (ER), we used a Kruskall-Wallis ranks test to compare identification accuracy rates across systems[3]. The main effect of SYSTEM was significant on RT ($F(14, 2219) = 2.58$, $p < .01$), and IT ($F(14, 2219) = 1.90, p < .05$). No significant main effect was found on ER ($\chi^2 = 13.88, p > .4$).[4]

Systems also differed significantly on Dice ($F(14, 2219) = 18.66, p < .001$) and MASI scores ($F(14, 2219) = 15.94, p < .001$).

In the remainder of this section, we first consider the differences between systems based on the algorithmic properties listed in Table 1 (and described in Section 1). For this part of the analysis, we consider intrinsic and extrinsic evaluation measures separately (Section 4.1 and 4.2). Subsequently, we explicitly compare the intrinsic and extrinsic methods (Section 4.3).

[3]The large number of zero values in ER proportions, and a high dependency of variance on the mean, meant that a non-parametric test was more appropriate.

[4]We left out the duplicated items in this analysis.

4.1 Differences between system types

As Table 1 indicates, there are some similarities between the new systems submitted for the ASGRE tasks, and the classic approaches discussed in Section 1. Eleven of the systems are incremental, although only a minority (5 systems) explicitly hardwire human preferences. Furthermore, 8 systems (all variations on two basic systems, CAM and TITCH) compute the discriminatory value of the properties selected. Just one system adopts the Full Brevity approach of Dale (1989), while the majority (11 systems) adopt the Dale and Reiter convention of always adding TYPE. Perhaps the greatest point of divergence between the ASGRE Challenge systems and the classic algorithms discussed in Section 1 is the predominance of data-driven approaches (12 systems are trainable), a characteristic that mirrors trends elsewhere in HLT.

It is worth asking how the particular algorithmic properties impact on performance, as measured with the evaluation methods used. Table 3 displays means computed with all the extrinsic and intrinsic measures, for all system properties except full brevity and discriminatory power. Since the latter are related to Minimality, which was a separate evaluation measure in this study, they are treated separately below (Section 4.2). Note that the aggregate scores in Table 3 are based on unequal-sized samples, since this is a post-hoc analysis which the evaluation studies described above did not directly target.

Table 3 indicates that systems that were directly trained on data display higher agreement (Dice and MASI) scores with the human data. This is relatively unsurprising; what is less straightforwardly predictable is that trainable systems have better task-performance scores (IT and RT). Systems which always include TYPE and those which are incremental apparently score better overall. Somewhat surprisingly, incorporating human attribute preferences results in smaller improvements than incrementality and TYPE inclusion. This is likely due to the fact that the attributes preferred by humans are not straightforwardly determinable in the people subdomain (van der Sluis et al., 2007). In contrast, those in the furniture domain, which include COLOUR, SIZE and ORIENTATION, are fairly predictable as far as human preference goes, based on previous psy-

	Minimality	Dice	MASI	RT	IT	ER
IS-FBN	1.35	0.770	0.601	1837.546	2188.923	6
DIT-DS	0	0.750	0.595	1304.119	1859.246	2
IS-IAC	0	0.746	0.597	1356.146	1973.193	6
CAM-T	0	0.725	0.560	1475.313	1978.237	5.33
CAM-TU	0	0.721	0.557	1297.372	1809.044	4
GRAPH-FP	4.73	0.689	0.479	1382.039	2053.326	3.33
GRAPH-SC	4.73	0.671	0.466	1349.047	1899.585	2
TITCH-RS+	0	0.669	0.459	1278.008	1814.933	1.33
TITCH-AS+	0	0.660	0.452	1321.204	1817.303	4.67
TITCH-RS	0	0.655	0.432	1255.278	1866.935	4.67
TITCH-AS	0	0.645	0.422	1229.417	1766.350	4.67
CAM-BU	10.14	0.630	0.420	1251.316	1877.948	4
NIL	20.27	0.625	0.477	1482.665	1960.314	5.33
CAM-B	8.11	0.620	0.403	1309.070	1952.394	5.33
IS-FBS	100	0.368	0.182	1461.453	2181.883	7.33

Table 2: Results for systems and evaluation measures (in order of Dice).

	trainable		includes type		incremental		human preferences	
	no	yes	no	yes	no	yes	no	yes
Dice	0.561	0.700	0.627	0.676	0.625	0.677	0.656	0.677
MASI	0.370	0.511	0.464	0.477	0.432	0.488	0.468	0.484
RT	1376.13	1371.41	1534.45	1313.83	1507.52	1323.63	1387.49	1343.02
IT	1993.13	1911.55	2076.08	1881.39	2080.93	1879.63	1932.87	1934.19
ER	5.2%	4.2%	6.3%	3.8%	4.7%	4.4%	4.5%	4.5%

Table 3: Means on intrinsic and extrinsic measures, by system type.

	Dice	MASI	IT	RT
Trainable	171.67*	145.63*	.002	4.36
Incremental	47.90*	45.61*	7.43	1.64
Includes type	18.42*	29.38*	6.89	11.27
Human pref.	3.27	5.12	2.11	.83
Incr. × Train.	40.82*	30.99*	.001	7.08

Table 4: Multivariate tests examining impact of system type on evaluation measures. Cells indicate $F-$values with 4 numerator and 2201 error degrees of freedom. * : $p \leq .05$ after Bonferroni correction.

cholinguistic results (Pechmann, 1989; Arts, 2004).

As a further test of these differences, a multivariate ANOVA was conducted, comparing scores on the intrinsic and extrinsic measures, as a function of the presence or absence of the 4 algorithmic properties being considered here. The results are shown in Table 4 which displays figures for all main effects and for the one interaction that was significant (incrementality × trainability, bottom row). Since this is a multiple-test post-hoc analysis on unequal-sized samples, all significance values are based on a Bonferroni correction[5] The table does not include ER, as no main effect of SYSTEM was found on this measure.

The results show that trainability, followed by incrementality and inclusion of TYPE had the strongest impact on system quality, as far as this is measured by Dice and MASI.

The incrementality × trainability interaction is mostly due to the huge effect that trainability has on how non-incremental systems perform. In the case of incremental systems, trainability gives rise to marginally better performance: the mean Dice score for non-trainable incremental systems was .625, as compared to .696 for trainable incremental systems. Similar patterns are observable with MASI (.433 vs .439). However, the most significant feature of the interaction is the large benefit that training on data confers on non-incremental systems. Untrained non-incremental systems obtained a mean Dice score of .368 (MASI = .182), while the mean Dice score on trained non-incremental systems was .710, with a mean MASI score of .515. Another,

[5]The Bonferroni correction adjusts for the increased likelihood of finding results statistically significant when multiple tests are applied by lowering the alpha value (the significance threshold) by dividing it by the number of tests involved.

Intrinsic measures		Extrinsic measures		
Dice	MASI	ER	RT	IT
-.904**	-.789**	.505	.183	.560*

Table 5: Correlations between Minimality and other evaluation measures. Legend: **: $p \leq .01$, *: $p \leq .05$

more tentative, conclusion that could be made by looking at these figures is that trainable systems perform better if they are non-incremental; however, the difference (Dice of .710 vs .696) seems rather marginal and would require further testing. What is also striking in this table is the absence of any significant impact of these properties on the task-performance measures of reading and identification time, despite initial impressions based on means (Table 3).

In summary, at least one of the principles that has been widely adopted in ASGRE – incrementality – seems to be validated by the evaluation results for these new algorithms. However, results show that improvement on an automatically computed humanlikeness metric such as MASI or Dice does not necessarily imply an improvement on a task-performance measure. This is a theme to which we return in Section 4.3.

4.2 The role of minimality

Apart from incrementality, the other dominant principle that has emerged from nearly three decades of ASGRE research is brevity. While psycholinguistic research has shown (Pechmann, 1989; Belke and Meyer, 2002; Arts, 2004) that the strict interpretation of the Gricean Quantity Maxim, adopted for example by Dale (1989), is not observed by speakers, brevity has remained a central concern in recent approaches (Gardent, 2002). Only one algorithm adopted a full-brevity approach in the ASGRE Challenge; however, several algorithms emphasised discriminatory power, and some of these, as a comparison between Tables 1 and 2 shows, tended to have higher Minimality scores overall.

Since Minimality was part of the overall design of the evaluation, it is possible to see the extent to which it covaries significantly with other performance measures. The relevant correlations are displayed in Table 5. There are highly significant *negative* correlations between Minimality and the two intrinsic humanlikeness measures (Dice and MASI). The significant positive correlation with IT implies that participants in our evaluation experiment tended to take longer to identify referents in the case of systems which produce more minimal descriptions.

The negative correlation with the humanlikeness measures corroborates previous findings that people do not observe a strict interpretation of the Gricean Quantity Maxim (Pechmann, 1989; Belke and Meyer, 2002; Arts, 2004). On the other hand, the direction of the correlation with IT is somewhat more surprising. One possible explanation is that minimal descriptions often do not include a TYPE attribute, since this seldom has any discriminatory power in the TUNA Corpus domains.

The role of Minimality remains a topic of some debate in the psycholinguistic literature. A recent study (Engelhardt et al., 2006) showed that identification of objects in a visual domain could be delayed if a description of the target referent was overspecified, suggesting a beneficial effect from Minimality. However, Engelhardt et al. also found, in a separate study, that overspecified descriptions tended to be rated as no worse than brief, or minimal ones. Indeed, one of the points that emerges from this study is that intrinsic and extrinsic assessments can yield contrasting results. It is to this issue that we now turn.

4.3 Intrinsic vs. extrinsic methods

Since humanlikeness (intrinsic) and task performance (extrinsic) are different perspectives on the outputs of an ASGRE algorithm, it is worth asking to what extent they agree. This is an important question for comparative HLT evaluation more generally, which has become dominated by corpus-based humanlikeness metrics. To obtain an indication of agreement, we computed correlations between the two humanlikeness measures and the task-performance measures; these are displayed in Table 6.

The two time measures (RT and IT) are strongly correlated, implying that when subjects took longer to read descriptions, they also took longer to identify a referent. This may seem surprising in view of the positive correlation between Minimality and IT (shorter descriptions imply *longer* IT) which was reported in Section 4.2, where Minimality was also

	RT	IT	ER	Dice	MASI
RT	1	.8**	0.46	0.12	.23
IT	.8**	1	.59*	-0.28	-.17
ER	0.46	.59*	1	-0.39	-.29
Dice	0.12	-0.28	-0.39	1	.97**
MASI	0.23	-0.17	-0.29	.97**	1

Table 6: Pairwise correlations between humanlikeness and task-performance measures (*: $p \leq .05$; **: $p \leq .01$)

shown not to correlate with RT (shorter descriptions do not imply longer or shorter RT). What this result indicates is that rather than the number of attributes in a description, as measured by Minimality, it is the content that influences identification latency. Part of the effect may be due to the lack of TYPE in minimal descriptions noted earlier.

Table 6 also shows that error rate is significantly correlated with IT (but not with RT), i.e. where subjects took longer to identify referents, the identification was more likely to be wrong. This points to a common underlying cause for slower reading and identification, possibly arising from the use of attributes (such as SIZE) which impose a greater cognitive load on the reader.

The very strong correlation (0.97) between Dice and MASI is to be expected, given the similarity in the way they are defined.

Another unambiguous result emerges: none of the similarity-based metrics covary significantly with any of the task-performance measures. An extended analysis involving a larger range of intrinsic metrics confirmed this lack of significant covariation for string-based similarity metrics as well as set-similarity metrics across two task-performance experiments (Belz and Gatt, 2008). This indicates that at least for some areas of HLT, task-performance evaluation is vital: without the external reality check provided by extrinsic evaluations, intrinsic evaluations may end up being too self-contained and disconnected from notions of usefulness to provide a meaningful assessment of systems' quality.

5 Conclusion

Comparative evaluation can be of great benefit, especially in an area as mature and diverse as GRE. A shared-task evaluation like the ASGRE Challenge can help identify the strengths and weaknesses of alternative approaches and techniques, as measured by different evaluation criteria. Perhaps even more importantly, such a challenge helps evaluate the evaluation methods themselves and reveals which (combinations of) evaluation methods are appropriate for a given evaluation purpose.

As far as the first issue is concerned, this paper has shown that trainability, incrementality and not aiming for minimality are the algorithmic properties most helpful in achieving high humanlikeness scores. This result is in line with psycholinguistic research on attribute selection in reference by humans. Less clear-cut is the relationship between these properties and task-performance measures that assess how efficiently a referent can be identified based on a description.

The second part of the analysis presented here showed that intrinsic and extrinsic perspectives on the quality of system outputs quality can yield uncorrelated sets of results, making it difficult to predict quality as measured by one based on quality as measured by the other. Furthermore, while intuitively it might be expected that higher humanlikeness entailed better task performance, our results suggest that this is not necessarily the case.

Our main conclusions for ASGRE evaluation are that (a) while humanlikeness evaluation may provide a measure of one aspect of systems but we need to be cautious in relying on humanlikeness as a criterion of overall quality (standard in evaluation of MT and summarisation); and (b) that we must not leave the NLG tradition of task-performance evaluations behind as we move towards more comparative forms of evaluation.

Acknowledgements

We gratefully acknowledge the contribution made to the evaluations by the faculty and staff at Brighton University who participated in the identification experiments. The biggest contribution was, of course, made by the participants in the ASGRE Challenge who created the systems involved in the evaluations: Bernd Bohnet, Ann Copestake, Pablo Gervás, Raquel Hervás, John Kelleher, Emiel Krahmer, Takahiro Kurosawa, Advaith Siddharthan, Philipp Spanger, Takenobu Tokunaga, Mariet Theune, Pascal Touset and Jette Viethen.

References

D. Appelt and A. Kronfeld. 1987. A computational model of referring. In *Proc. 10th International Joint Conference on Artificial Intelligence (IJCAI-87)*, pages 640–647.

D. Appelt. 1985. Planning English referring expressions. *Artificial Intelligence*, 26(1):1–33.

A. Arts. 2004. *Overspecification in Instructive Texts*. Ph.D. thesis, University of Tilburg.

E. Belke and A. Meyer. 2002. Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing times during same-different decisions. *European Journal of Cognitive Psychology*, 14(2):237–266.

A. Belz and A. Gatt. 2007. The Attribute Selection for GRE Challenge: Overview and evaluation results. In *Proc. 2nd UCNLG Workshop: Language Generation and Machine Translation (UCNLG+MT)*, pages 75–83.

A. Belz and A. Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proc. 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*. To appear.

R. Dale and N. Haddock. 1991. Generating referring expressions containing relations. In *Proc. 5th Conference of the European Chapter of the Association for Computational Linguistics (EACL-91).*, pages 161–166.

R. Dale and E. Reiter. 1995. Computational interpretation of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(8):233–263.

R. Dale and M. White, editors. 2007. *Shared tasks and comparative evaluation in Natural Language Generation: Workshop Report*.

Robert Dale. 1989. Cooking up referring expressions. In *Proc. 27th Annual Meeting of the Association for Computational Linguistics (ACL-89).*, pages 68–75.

P. E. Engelhardt, K.G.D Bailey, and F. Ferreira. 2006. Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, 54:554–573.

K. I. Forster and J. C. Forster. 2003. DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35(1):116–124.

C. Gardent. 2002. Generating minimal definite descriptions. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL-02).*, pages 96–103.

A. Gatt and K. van Deemter. 2007. Incremental generation of plural descriptions: Similarity and partitioning. In *Proc. 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CONLL-07)*, pages 102–111.

A. Gatt, I. van der Sluis, and K. van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proc. 11th European Workshop on Natural Language Generation (ENLG-07).*, pages 49–56.

H. P. Grice. 1975. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics: Speech Acts.*, volume III. Academic Press.

S. Gupta and A. J. Stent. 2005. Automatic evaluation of referring expression generation using corpora. In *Proc. 1st Workshop on Using Corpora in NLG (UCNLG-05).*, pages 1–6.

H. Horacek. 2004. On referring to sets of objects naturally. In *Proc. 3rd International Conference on Natural Language Generation (INLG-04).*, pages 70–79.

A. Karasimos and A. Isard. 2004. Multilingual evaluation of a natural language generation system. In *Proc. 4th International Conference on Language Resources and Evaluation (LREC-04).*

J. D. Kelleher and G-J Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proc. joint 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL/COLING-06).*, pages 1041–1048.

R. Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proc. 5th International Conference on Language Resources and Evaluation (LREC-06).*, pages 831–836.

Thomas Pechmann. 1989. Incremental speech production and referential overspecification. *Linguistics*, 27:89–110.

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

E. Reiter. 1990. The computational complexity of avoiding conversational implicatures. In *Proc. 28th Annual Meeting of the Association for Computational Linguistics (ACL-90).*, pages 97–104.

K. van Deemter. 2002. Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics*, 28(1):37–52.

K. van Deemter. 2006. Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222.

I. van der Sluis, A. Gatt, and K. van Deemter. 2007. Evaluating algorithms for the generation of referring expressions: Going beyond toy domains. In *Proc. International Conference on Recent Advances in Natural Language Processing (RANLP-07).*

J. Viethen and R. Dale. 2006. Algorithms for generating referring expressions: Do they do what people do? In *Proc. 4th International Conference on Natural Language Generation (INLG-06).*, pages 63–72.