Intrinsic vs. Extrinsic Evaluation Measures for Referring Expression Generation

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

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

Abstract

In this paper we present research in which we apply (i) the kind of intrinsic evaluation metrics that are characteristic of current comparative HLT evaluation, and (ii) extrinsic, human task-performance evaluations more in keeping with NLG traditions, to 15 systems implementing a language generation task. We analyse the evaluation results and find that there are no significant correlations between intrinsic and extrinsic evaluation measures for this task.

1 Introduction

In recent years, NLG evaluation has taken on a more comparative character. NLG now has evaluation results for comparable, but independently developed systems, including results for systems that regenerate the Penn Treebank (Langkilde, 2002) and systems that generate weather forecasts (Belz and Reiter, 2006). The growing interest in comparative evaluation has also resulted in a tentative interest in shared-task evaluation events, which led to the first such event for NLG (the Attribute Selection for Generation of Referring Expressions, or ASGRE, Challenge) in 2007 (Belz and Gatt, 2007), with a second event (the Referring Expression Generation, or REG, Challenge) currently underway.

In HLT in general, comparative evaluations (and shared-task evaluation events in particular) are dominated by intrinsic evaluation methodologies, in contrast to the more extrinsic evaluation traditions of NLG. In this paper, we present research in which we applied both intrinsic and extrinsic evaluation methods to the same task, in order to shed light on how the two correlate for NLG tasks. The results show a surprising lack of correlation between the two types of measures, suggesting that intrinsic metrics and extrinsic methods can represent two very different views of how well a system performs.

2 Task, Data and Systems

Referring expression generation (REG) is concerned with the generation of expressions that describe entities in a given piece of discourse. REG research goes back at least to the 1980s (Appelt, Grosz, Joshi, McDonald and others), but the field as it is today was shaped in particular by Dale and Reiter's work (Dale, 1989; Dale and Reiter, 1995). REG tends to be divided into the stages of *attribute selection* (selecting properties of entities) and *realisation* (converting selected properties into word strings). Attribute selection in its standard formulation was the shared task in the ASGRE Challenge: given an intended referent ('target') and the other domain entities ('distractors') each with possible attributes, select a set of attributes for the target referent.

The ASGRE data (which is now publicly available) consists of all 780 singular items in the TUNA corpus (Gatt et al., 2007) in two subdomains, consisting of descriptions of furniture and people. Each data item is a paired attribute set (as derived from a human-produced RE) and domain representation (target and distractor entities represented as possible attributes and values).

ASGRE participants were asked to submit the outputs produced by their systems for an unseen test data set. The outputs from 15 of these systems, shown in the left column of Table 1, were used in

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 197–200,
Columbus, Ohio, USA, June 2008.

the experiments reported below. Systems differed in terms of whether they were trainable, performed exhaustive search and hardwired use of certain attributes types, among other algorithmic properties (see the ASGRE papers for full details). In the case of one system (IS-FBS), a buggy version was originally submitted and used in Exp 1. It was replaced in Exp 2 by a corrected version; the former is marked by a * in what follows.

3 Evaluation Methods

1. Extrinsic evaluation measures: We conducted two task-performance evaluation experiments (the first was part of the ASGRE Challenge, the second is new), in which participants identified the referent denoted by a description by clicking on a picture in a visual display of target and distractor entities. To enable subjects to read the outputs of peer systems, we converted them from the attribute-value format described above to something more readable, using a simple attribute-to-word converter.

Both experiments used a Repeated Latin Squares design, and involved 30 participants and 2,250 individual trials (see Belz & Gatt (2007) for full details).

In Exp 1, subjects were shown the domain on the same screen as the description. Two dependent measures were used: (i) combined reading and identification time (RIT), measured from the point at which the description and pictures appeared on the screen to the point at which a picture was selected by mouse-click; and (ii) error rate (ER-1).

In Exp 2, subjects first read the description and then initiated the presentation of domain entities. We computed: (i) reading time (RT), measured from the presentation of a description to the point where a subject requested the presentation of the domain; (ii) identification time (IT), measured from the presentation of the domain to the point where a subject clicked on a picture; and (iii) error rate (ER-2).

2. REG-specific intrinsic measures: *Uniqueness* is the proportion of attribute sets generated by a system which identify the referent uniquely (i.e. none of the distractors). *Minimality* is the proportion of attribute sets which are minimal as well as unique (i.e. there is no smaller unique set of attributes). These measures were included because they are commonly named as desiderata for attribute selection algorithms in the REG field (Dale, 1989). The minimality check used in this paper treats referent type as a simple attribute, as the ASGRE systems tended to do.[1]

3. Set-similarity measures: The *Dice similarity coefficient* computes the similarity between a peer attribute set A_1 and a (human-produced) reference attribute set A_2 as $\frac{2\times|A_1\cap A_2|}{|A_1|+|A_2|}$. *MASI* (Passonneau, 2006) is similar but biased in favour of similarity where one set is a subset of the other.

4. String-similarity measures: In order to apply string-similarity metrics, peer and reference outputs were converted to word-strings by the method described under 1 above. *String-edit distance* (SE) is straightforward Levenshtein distance with a substitution cost of 2 and insertion/deletion cost of 1. We also used the version of string-edit distance ('SEB') of Bangalore et al. (2000) which normalises for length. *BLEU* computes the proportion of word n-grams ($n \leq 4$ is standard) that a peer output shares with several reference outputs. The *NIST* MT evaluation metric (Doddington, 2002) is an adaptation of BLEU which gives more importance to less frequent (hence more informative) n-grams. We also used two versions of the ROUGE metric (Lin and Hovy, 2003), *ROUGE-2* and *ROUGE-SU4* (based on non-contiguous, or 'skip', n-grams), which were official scores in the DUC 2005 summarization task.

4 Results

Results for all evaluation measures and all systems are shown in Table 1. Uniqueness results are not included, as all systems scored 100%.

We ran univariate analyses of variance (ANOVAs) using SYSTEM as the independent variable (15 levels), testing its effect on the extrinsic task-performance measures. For error rate (ER), we used a Kruskal-Wallis ranks test to compare identification accuracy rates across systems[2]. The main effect of SYSTEM was significant on RIT ($F(14, 2249) = 6.401$, $p < .001$), RT ($F(14, 2249) = 2.56$, $p < .01$), and IT ($F(14, 2249) = 1.93$, $p < .01$). In neither experiment was there a significant effect on ER.

[1] As a consequence, the Minimality results we report here look different from those in the ASGRE report.

[2] A non-paramteric test was more appropriate given the large number of zero values in ER proportions, and a high dependency of variance on the mean.

	extrinsic					*REG*	*string-similarity*						*set-similarity*	
	RIT	RT	IT	ER-1	ER-2	Min	RSU4	R-2	NIST	BLEU	SE	SEB	Dice	MASI
CAM-B	2784.80	1309.07	1952.39	9.33	5.33	8.11	.673	.647	2.70	.309	4.42	.307	.620	.403
CAM-BU	2659.37	1251.32	1877.95	9.33	4	10.14	.663	.638	2.61	.317	4.23	.359	.630	.420
CAM-T	2626.02	1475.31	1978.24	10	5.33	0	.698	.723	3.50	.415	3.67	.496	.725	.560
CAM-TU	2572.82	1297.37	1809.04	8.67	4	0	.677	.691	3.28	.407	3.71	.494	.721	.557
DIT-DS	2785.40	1304.12	1859.25	10.67	2	0	.651	.679	4.23	.457	3.55	.525	.750	.595
GR-FP	2724.56	1382.04	2053.33	8.67	3.33	4.73	.65	.649	3.24	.358	3.87	.441	.689	.480
GR-SC	2811.09	1349.05	1899.59	11.33	2	4.73	.644	.644	2.42	.305	4	.431	.671	.466
IS-FBN	3570.90	1837.55	2188.92	15.33	6	1.35	.771	.772	4.75	.521	3.15	.438	.770	.601
IS-FBS	–	1461.45	2181.88	–	7.33	100	.485	.448	2.11	.166	5.53	.089	.368	.182
*IS-FBS	4008.99	–	–	10	–	39.86	–	–	–	–	–	–	.527	.281
IS-IAC	2844.17	1356.15	1973.19	8.67	6	0	.612	.623	3.77	.442	3.43	.559	.746	.597
NIL	1960.31	1482.67	1960.31	10	5.33	20.27	.525	.509	3.32	.32	4.12	.447	.625	.477
T-AS+	2652.85	1321.20	1817.30	9.33	4.67	0	.671	.684	2.62	.298	4.24	.37	.660	.452
T-AS	2864.93	1229.42	1766.35	10	4.67	0	.683	.692	2.99	.342	4.10	.393	.645	.422
T-RS+	2759.76	1278.01	1814.93	6.67	1.33	0	.677	.697	2.85	.303	4.32	.36	.669	.459
T-RS	2514.37	1255.28	1866.94	8.67	4.67	0	.694	.711	3.16	.341	4.18	.383	.655	.432

Table 1: Results for all systems and evaluation measures (ER-1 = error rate in Exp 1, ER-2 = error rate in Exp 2). (R = ROUGE; system IDs as in the ASGRE papers, except GR = GRAPH; T = TITCH).

Table 2 shows correlations between the automatic metrics and the task-performance measures from Exp 1. RIT and ER-1 are not included because of the presence of *IS-FBS in Exp 1 (but see individual results below). For reasons of space, we refer the reader to the table for individual correlation results.

We also computed correlations between the task-performance measures across the two experiments (leaving out the IS-FBS system). Correlation between RIT and RT was .827**; between RIT and IT .675**; and there was no significant correlation between the error rates. The one difference evident between RT and IT is that ER correlates only with IT (not RT) in Exp 2 (see Table 2).

5 Discussion

In Table 2, the four broad types of metrics we have investigated (task-performance, REG-specific, string similarity, set similarity) are indicated by vertical and horizontal lines. The results within each of the resulting boxes are very homogeneous. There are significant (and mostly strong) correlations not only among the string-similarity metrics and among the set-similarities, but also across the two types. There are also significant correlations between the three task-performance measures.

However, the correlation figures between the task-performance measures and all others are weak and not significant. The one exception is the correlation between NIST and RT which is actually in the wrong direction (better NIST implies *worse* reading times).

This is an unambiguous result and it shows clearly that similarity to human-produced reference texts is not necessarily indicative of quality as measured by human task performance.

The emergence of comparative evaluation in NLG raises the broader question of how systems that generate language should be compared. In MT and summarisation it is more or less taken as read that systems which generate more human-like language are better systems. However, it has not been shown that more human-like outputs result in better performance from an extrinsic perspective. Intuitively, it might be expected that higher humanlikeness entails better task-performance (here, shorter reading/identification times, lower error). The lack of significant covariation between intrinsic and extrinsic measures in our experiments suggests otherwise.

6 Conclusions

Our aim in this paper was to shed light on how the intrinsic evaluation methodologies that dominate current comparative HLT evaluations correlate with human task-performance evaluations more in keeping with NLG traditions. We used the data and systems from the recent ASGRE Challenge, and compared a total of 17 different evaluation methods for 15 different systems implementing the ASGRE task.

Our most striking result is that none of the metrics that assess humanlikeness correlate with any of the task-performance measures, while strong correlations are observed *within* the two classes of mea-

	extrinsic			*REG*	*string-similarity*						*set-similarity*	
	RT	IT	ER-2	Min	R-SU4	R-2	NIST	BLEU	SE	SEB	Dice	MASI
RT	1	.8**	.46	.18	.10	.05	.54*	.39	-.30	.02	.12	.23
IT	.8**	1	.59*	.56*	-.24	-.33	.22	.04	.09	-.31	-.28	-.17
ER-2	.46	.59*	1	.51	-.29	-.36	.03	-.08	.22	-.34	-.39	-.29
Min	.18	.56*	.51	1	-.76**	-.81**	-.46	-.66**	.79**	-.8**	-.90**	-.79**
R-SU4	.10	-.24	-.29	-.76**	1	.98**	.45	.63*	-.63*	.42	.72**	.57*
R-2	.05	-.33	-.36	-.81**	.98**	1	.51	.68**	-.69**	.53*	.78**	.65**
NIST	.54*	.22	.03	-.46	.45	.51	1	.94**	-.84**	.68**	.74**	.82**
BLEU	.39	.04	-.08	-.66**	.63*	.68**	.94**	1	-.96**	.82**	.89**	.93**
SE	-.30	.09	.22	.79**	-.63*	-.69**	-.84**	-.96**	1	-.92**	-.96**	-.97**
SEB	.02	-.31	-.34	-.8**	.42	.53*	.68**	.82**	-.92**	1	.92**	.95**
Dice	.12	-.28	-.39	-.90**	.72**	.78**	.74**	.89**	-.96**	.92**	1	.97**
MASI	.23	-.17	-.29	-.79**	.57*	.65**	.82**	.93**	-.97**	.95**	.97**	1

Table 2: Pairwise correlations between all automatic measures and the task-performance results from Exp 2. (* = significant at .05; ** at .01). R = ROUGE.

sures – intrinsic and extrinsic. Somewhat worryingly, our results show that a system's ability to produce human-like outputs may be completely unrelated to its effect on human task-performance.

Our main conclusions for REG evaluation are that we need to be cautious in relying on humanlikeness as a quality criterion, and that we leave extrinsic evaluation behind at our peril as we move towards more comparative forms of evaluation.

Given that the intrinsic metrics that dominate in competetive HLT evaluations are not assessed in terms of correlation with extrinsic notions of quality, our results sound a more general note of caution about using intrinsic measures (and humanlikeness metrics in particular) without extrinsic validation.

Acknowledgments

We gratefully acknowledge the contribution made to the evaluations by the faculty and staff at Brighton University who participated in the identification experiments. Thanks are also due to Robert Dale, Kees van Deemter, Ielka van der Sluis and the anonymous reviewers for very helpful comments. The biggest contribution was, of course, made by the participants in the ASGRE Challenge who created the systems involved in the evaluations.

References

S. Bangalore, O. Rambow, and S. Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the 1st International Conference on Natural Language Generation (INLG '00)*, pages 1–8.

A. Belz and A. Gatt. 2007. The attribute selection for GRE challenge: Overview and evaluation results. In *Proceedings of the 2nd UCNLG Workshop: Language Generation and Machine Translation (UCNLG+MT)*, pages 75–83.

A. Belz and E. Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proc. EACL'06*, pages 313–320.

R. Dale and E. Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

R. Dale. 1989. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. ARPA Workshop on Human Language Technology*.

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

I. Langkilde. 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the 2nd International Natural Language Generation Conference (INLG '02)*.

C.-Y. Lin and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. HLT-NAACL 2003*, pages 71–78.

R. Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the 5th Language Resources and Evaluation Converence (LREC'06)*.