Why We Need New Evaluation Metrics for NLG

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry and **Verena Rieser**
School of Mathematical and Computer Sciences
Heriot-Watt University, Edinburgh
`j.novikova, o.dusek, ac293, v.t.rieser@hw.ac.uk`

Abstract

The majority of NLG evaluation relies on automatic metrics, such as BLEU. In this paper, we motivate the need for novel, system- and data-independent automatic evaluation methods: We investigate a wide range of metrics, including state-of-the-art word-based and novel grammar-based ones, and demonstrate that they only weakly reflect human judgements of system outputs as generated by data-driven, end-to-end NLG. We also show that metric performance is data- and system-specific. Nevertheless, our results also suggest that automatic metrics perform reliably at system-level and can support system development by finding cases where a system performs poorly.

1 Introduction

Automatic evaluation measures, such as BLEU (Papineni et al., 2002), are used with increasing frequency to evaluate Natural Language Generation (NLG) systems: Up to 60% of NLG research published between 2012–2015 relies on automatic metrics (Gkatzia and Mahamood, 2015). Automatic evaluation is popular because it is cheaper and faster to run than human evaluation, and it is needed for automatic benchmarking and tuning of algorithms. The use of such metrics is, however, only sensible if they are known to be sufficiently correlated with human preferences. This is rarely the case, as shown by various studies in NLG (Stent et al., 2005; Belz and Reiter, 2006; Reiter and Belz, 2009), as well as in related fields, such as dialogue systems (Liu et al., 2016), machine translation (MT) (Callison-Burch et al., 2006), and image captioning (Elliott and Keller, 2014; Kilickaya et al., 2017). This paper follows on from the above previous work and presents another evaluation study into automatic metrics with the aim to firmly establish the need for new metrics. We consider this paper to be the most complete study to date, across metrics, systems, datasets and domains, focusing on recent advances in data-driven NLG. In contrast to previous work, we are the first to:

- Target end-to-end data-driven NLG, where we compare 3 different approaches. In contrast to NLG methods evaluated in previous work, our systems can produce ungrammatical output by (a) generating word-by-word, and (b) learning from noisy data.
- Compare a large number of 21 automated metrics, including novel grammar-based ones.
- Report results on two different domains and three different datasets, which allows us to draw more general conclusions.
- Conduct a detailed error analysis, which suggests that, while metrics can be reasonable indicators at the system-level, they are not reliable at the sentence-level.
- Make all associated code and data publicly available, including detailed analysis results.[1]

2 End-to-End NLG Systems

In this paper, we focus on recent end-to-end, data-driven NLG methods, which jointly learn sentence planning and surface realisation from non-aligned data (Dušek and Jurčíček, 2015; Wen et al., 2015; Mei et al., 2016; Wen et al., 2016; Sharma et al., 2016; Dušek and Jurčíček, 2016, Lampouras and Vlachos, 2016). These approaches do not require costly semantic alignment between Meaning Representations (MR) and human references (also referred to as "ground truth" or "targets"), but are

[1] Available for download at: `https://github.com/jeknov/EMNLP_17_submission`

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2241–2252
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

System	Dataset			Total
	BAGEL	SFREST	SFHOTEL	
LOLS	202	581	398	1,181
RNNLG	-	600	477	1,077
TGEN	202	-	-	202
Total	404	1,181	875	2,460

Table 1: Number of NLG system outputs from different datasets and systems used in this study.

based on parallel datasets, which can be collected in sufficient quality and quantity using effective crowdsourcing techniques, e.g. (Novikova et al., 2016), and as such, enable rapid development of NLG components in new domains. In particular, we compare the performance of the following systems:

- **RNNLG:**[2] The system by Wen et al. (2015) uses a Long Short-term Memory (LSTM) network to jointly address sentence planning and surface realisation. It augments each LSTM cell with a gate that conditions it on the input MR, which allows it to keep track of MR contents generated so far.
- **TGEN:**[3] The system by Dušek and Jurčíček (2015) learns to incrementally generate deep-syntax dependency trees of candidate sentence plans (i.e. which MR elements to mention and the overall sentence structure). Surface realisation is performed using a separate, domain-independent rule-based module.
- **LOLS:**[4] The system by Lampouras and Vlachos (2016) learns sentence planning and surface realisation using Locally Optimal Learning to Search (LOLS), an imitation learning framework which learns using BLEU and ROUGE as non-decomposable loss functions.

3 Datasets

We consider the following crowdsourced datasets, which target utterance generation for spoken dialogue systems. Table 1 shows the number of system outputs for each dataset. Each data instance consists of one MR and one or more natural language references as produced by humans, such as the following example, taken from the BAGEL dataset:[5]

> **MR:** inform(name=X, area=X, pricerange=moderate, type=restaurant)
> **Reference:** *"X is a moderately priced restaurant in X."*

- **SFHOTEL & SFREST** (Wen et al., 2015) provide information about hotels and restaurants in San Francisco. There are 8 system dialogue act types, such as *inform*, *confirm*, *goodbye* etc. Each domain contains 12 attributes, where some are common to both domains, such as *name, type, pricerange, address, area,* etc., and the others are domain-specific, e.g. *food* and *kids-allowed* for restaurants; *hasinternet* and *dogs-allowed* for hotels. For each domain, around 5K human references were collected with 2.3K unique human utterances for SFHOTEL and 1.6K for SFREST. The number of unique system outputs produced is 1181 for SFREST and 875 for SFHOTEL.
- **BAGEL** (Mairesse et al., 2010) provides information about restaurants in Cambridge. The dataset contains 202 aligned pairs of MRs and 2 corresponding references each. The domain is a subset of SFREST, including only the *inform* act and 8 attributes.

4 Metrics

4.1 Word-based Metrics (WBMs)

NLG evaluation has borrowed a number of automatic metrics from related fields, such as MT, summarisation or image captioning, which compare output texts generated by systems to ground-truth references produced by humans. We refer to this group as word-based metrics. In general, the higher these scores are, the better or more similar to the human references the output is.[6] The following order reflects the degree these metrics move from simple n-gram overlap to also considering term frequency (TF-IDF) weighting and semantically similar words.

- **Word-overlap Metrics (WOMs):** We consider frequently used metrics, including TER (Snover et al., 2006), BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), NIST (Doddington, 2002), LEPOR (Han et al., 2012), CIDER (Vedantam et al., 2015), and METEOR (Lavie and Agarwal, 2007).
- **Semantic Similarity (SIM):** We calculate the Semantic Text Similarity measure designed by Han et al. (2013). This measure is based on distributional similarity and Latent Semantic Analysis

[2] https://github.com/shawnwun/RNNLG

[3] https://github.com/UFAL-DSG/tgen

[4] https://github.com/glampouras/JLOLS_NLG

[5] Note that we use lexicalised versions of SFHOTEL and SFREST and a partially lexicalised version of BAGEL, where proper names and place names are replaced by placeholders ("X"), in correspondence with the outputs generated by the systems, as provided by the system authors.

[6] Except for TER whose scale is reversed.

(LSA) and is further complemented with semantic relations extracted from WordNet.

4.2 Grammar-based metrics (GBMs)

Grammar-based measures have been explored in related fields, such as MT (Giménez and Màrquez, 2008) or grammatical error correction (Napoles et al., 2016), and, in contrast to WBMs, do not rely on ground-truth references. To our knowledge, we are the first to consider GBMs for sentence-level NLG evaluation. We focus on two important properties of texts here – readability and grammaticality:

- **Readability** quantifies the difficulty with which a reader understands a text, as used for e.g. evaluating summarisation (Kan et al., 2001) or text simplification (Francois and Bernhard, 2014). We measure readability by the Flesch Reading Ease score (**RE**) (Flesch, 1979), which calculates a ratio between the number of characters per sentence, the number of words per sentence, and the number of syllables per word. Higher RE score indicates a less complex utterance that is easier to read and understand. We also consider related measures, such as characters per utterance (**len**) and per word (**cpw**), words per sentence (**wps**), syllables per sentence (**sps**) and per word (**spw**), as well as polysyllabic words per utterance (**pol**) and per word (**ppw**). The higher these scores, the more complex the utterance.
- **Grammaticality:** In contrast to previous NLG methods, our corpus-based end-to-end systems can produce ungrammatical output by (a) generating word-by-word, and (b) learning from noisy data. As a first approximation of grammaticality, we measure the number of misspellings (**msp**) and the parsing score as returned by the Stanford parser (**prs**). The lower the **msp**, the more grammatically correct an utterance is. The Stanford parser score is not designed to measure grammaticality, however, it will generally prefer a grammatical parse to a non-grammatical one.[7] Thus, lower parser scores indicate less grammatically-correct utterances. In future work, we aim to use specifically designed grammar-scoring functions, e.g. (Napoles et al., 2016), once they become publicly available.

5 Human Data Collection

To collect human rankings, we presented the MR together with 2 utterances generated by different systems side-by-side to crowdworkers, which were asked to score each utterance on a 6-point Likert scale for:

- **Informativeness:** *Does the utterance provide all the useful information from the meaning representation?*
- **Naturalness:** *Could the utterance have been produced by a native speaker?*
- **Quality:** *How do you judge the overall quality of the utterance in terms of its grammatical correctness and fluency?*

Each system output (see Table 1) was scored by 3 different crowdworkers. To reduce participants' bias, the order of appearance of utterances produced by each system was randomised and crowdworkers were restricted to evaluate a maximum of 20 utterances. The crowdworkers were selected from English-speaking countries only, based on their IP addresses, and asked to confirm that English was their native language.

To assess the reliability of ratings, we calculated the intra-class correlation coefficient (ICC), which measures inter-observer reliability on ordinal data for more than two raters (Landis and Koch, 1977). The overall ICC across all three datasets is 0.45 ($p < 0.001$), which corresponds to a moderate agreement. In general, we find consistent differences in inter-annotator agreement per system and dataset, with lower agreements for LOLS than for RNNLG and TGEN. Agreement is highest for the SFHOTEL dataset, followed by SFREST and BAGEL (details provided in supplementary material).

6 System Evaluation

Table 2 summarises the individual systems' overall corpus-level performance in terms of automatic and human scores (details are provided in the supplementary material).

All WOMs produce similar results, with SIM showing different results for the restaurant domain (BAGEL and SFREST). Most GBMs show the same trend (with different levels of statistical significance), but RE is showing inverse results. System performance is dataset-specific: For WBMs, the LOLS system consistently produces better results on BAGEL compared to TGEN, while for SFREST and SFHOTEL, LOLS is outperformed by RNNLG in

[7] http://nlp.stanford.edu/software/parser-faq.shtml

	BAGEL		SFHOTEL		SFREST	
metric	TGEN	LOLS	RNNLG	LOLS	RNNLG	LOLS
WOMs		More overlap	More overlap*		More overlap*	
SIM	More similar		More similar*			More similar
GBMs	Better grammar(*)		Better grammar(*)		Better grammar	
RE		More complex*		More complex*		More complex*
inform	4.77(Sd=1.09)	**4.91**(Sd=1.23)	**5.47***(Sd=0.81)	5.27(Sd=1.02)	**5.29***(Sd=0.94)	5.16(Sd=1.07)
natural	**4.76**(Sd=1.26)	4.67(Sd=1.25)	**4.99***(Sd=1.13)	4.62(Sd=1.28)	**4.86** (Sd=1.13)	4.74(Sd=1.23)
quality	**4.77**(Sd=1.19)	4.54(Sd=1.28)	**4.54** (Sd=1.18)	4.53(Sd=1.26)	4.51 (Sd=1.14)	**4.58**(Sd=1.33)

Table 2: System performance per dataset (summarised over metrics), where "*" denotes $p < 0.05$ for all the metrics and "(*)" shows significance on $p < 0.05$ level for the majority of the metrics.

terms of WBMs. We observe that human *informativeness* ratings follow the same pattern as WBMs, while the average similarity score (SIM) seems to be related to human *quality* ratings.

Looking at GBMs, we observe that they seem to be related to *naturalness* and *quality* ratings. Less complex utterances, as measured by readability (RE) and word length (cpw), have higher *naturalness* ratings. More complex utterances, as measured in terms of their length (len), number of words (wps), syllables (sps, spw) and polysyllables (pol, ppw), have lower *quality* evaluation. Utterances measured as more grammatical are on average evaluated higher in terms of *naturalness*.

These initial results suggest a relation between automatic metrics and human ratings at system level. However, average scores can be misleading, as they do not identify worst-case scenarios. This leads us to inspect the correlation of human and automatic metrics for each MR-system output pair at utterance level.

7 Relation of Human and Automatic Metrics

7.1 Human Correlation Analysis

We calculate the correlation between automatic metrics and human ratings using the Spearman coefficient (ρ). We split the data per dataset and system in order to make valid pairwise comparisons. To handle outliers within human ratings, we use the median score of the three human raters.[8] Following Kilickaya et al. (2017), we use the Williams' test (Williams, 1959) to determine significant differences between correlations. Table 3 summarises the utterance-level correlation results between automatic metrics and human ratings, listing the best (i.e. highest absolute ρ) results for each type of metric (details provided in supplementary material). Our results suggest that:

- In sum, no metric produces an even moderate correlation with human ratings, independently of dataset, system, or aspect of human rating. This contrasts with our initially promising results on the system level (see Section 6) and will be further discussed in Section 8. Note that similar inconsistencies between document- and sentence-level evaluation results are observed in MT (Specia et al., 2010).
- Similar to our results in Section 6, we find that WBMs show better correlations to human ratings of *informativeness* (which reflects content selection), whereas GBMs show better correlations to *quality* and *naturalness*.
- Human ratings for *informativeness*, *naturalness* and *quality* are highly correlated with each other, with the highest correlation between the latter two ($\rho = 0.81$) reflecting that they both target surface realisation.
- All WBMs produce similar results (see Figure 1 and 2): They are strongly correlated with each other, and most of them produce correlations with human ratings which are *not* significantly different from each other. GBMs, on the other hand, show greater diversity.
- Correlation results are system- and dataset-specific (details provided in supplementary material). We observe the highest correlation for TGEN on BAGEL (Figures 1 and 2) and LOLS on SFREST, whereas RNNLG often shows low correlation between metrics and human ratings. This lets us conclude that WBMs and GBMs are sensitive to different systems and datasets.
- The highest positive correlation is observed between the number of words (wps) and *informative-*

[8] As an alternative to using the median human judgment for each item, a more effective way to use all the human judgments could be to use Hovy et al. (2013)'s MACE tool for inferring the reliability of judges.

		BAGEL		SFHOTEL		SFREST	
		TGEN	LOLS	RNNLG	LOLS	RNNLG	LOLS
Best WBM	inform.	0.30* (BLEU-1)	0.20* (ROUGE)	0.09 (BLEU-1)	0.14* (LEPOR)	0.13* (SIM)	0.28* (LEPOR)
	natural.	-0.19* (TER)	-0.19* (TER)	0.10* (METEOR)	-0.20* (TER)	0.17* (ROUGE)	0.19* (METEOR)
	quality	-0.16* (TER)	0.16* (METEOR)	0.10* (METEOR)	-0.12* (TER)	0.09* (METEOR)	0.18* (LEPOR)
Best GBM	inform.	0.33* (wps)	0.16* (ppw)	-0.09 (ppw)	0.13* (cpw)	0.11* (len)	0.21* (len)
	natural.	-0.25* (len)	-0.28* (wps)	-0.17* (len)	-0.18* (sps)	-0.19* (wps)	-0.21* (sps)
	quality	-0.19* (cpw)	0.31* (prs)	-0.16* (ppw)	-0.17* (spw)	0.11* (prs)	-0.16* (sps)

Table 3: Highest absolute Spearman correlation between metrics and human ratings, with "*" denoting $p < 0.05$ (metric with the highest absolute value of ρ given in brackets).

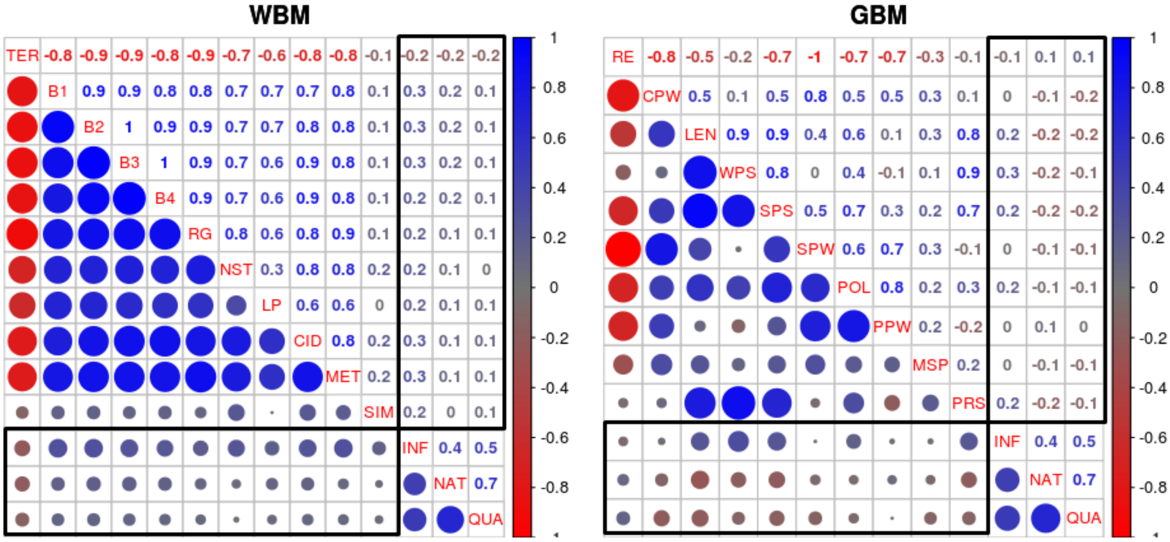

Figure 1: Spearman correlation results for TGEN on BAGEL. Bordered area shows correlations between human ratings and automatic metrics, the rest shows correlations among the metrics. Blue colour of circles indicates positive correlation, while red indicates negative correlation. The size of circles denotes the correlation strength.

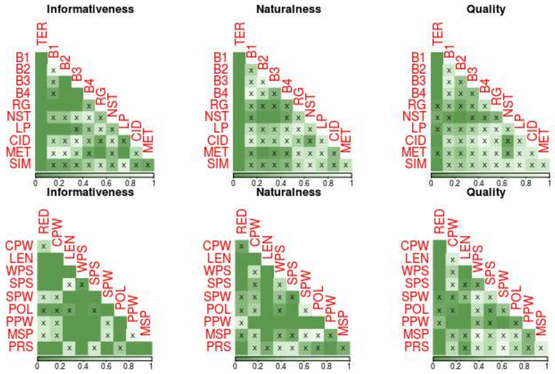

Figure 2: Williams test results: X represents a *non*-significant difference between correlations ($p < 0.05$; top: WBMs, bottom: GBMs).

ness for the TGEN system on BAGEL ($\rho = 0.33$, $p < 0.01$, see Figure 1). However, the wps metric (amongst most others) is not robust across systems and datasets: Its correlation on other datasets is very weak, ($\rho \leq .18$) and its correlation with informativeness ratings of LOLS outputs is insignificant.

- As a sanity check, we also measure a random score $[0.0, 1.0]$ which proves to have a close-to-zero correlation with human ratings (highest $\rho = 0.09$).

7.2 Accuracy of Relative Rankings

We now evaluate a more coarse measure, namely the metrics' ability to predict relative human ratings. That is, we compute the score of each metric for two system output sentences corresponding to the same MR. The prediction of a metric is correct if it orders the sentences in the same way as median human ratings (note that ties are allowed). Following previous work (Vedantam et al., 2015; Kilickaya et al., 2017), we mainly concentrate on WBMs. Results summarised in Table 4 show that most metrics' performance is not significantly different from that of a random score (Wilcoxon signed rank test). While the random score fluctuates between 25.4–44.5% prediction accuracy, the metrics achieve an accuracy of between 30.6–49.8%. Again, the performance of the metrics is dataset-specific: Metrics perform best on BAGEL data; for SFHOTEL, metrics show mixed performance while for SFREST, metrics perform worst.

		informat.	**naturalness**	**quality**
BAGEL	raw data	TER, BLEU1-4, ROUGE, NIST, LEPOR, CIDEr, METEOR, SIM	TER, BLEU1-4, ROUGE, NIST, LEPOR, CIDEr, METEOR, SIM	TER, BLEU1-4, ROUGE, NIST, LEPOR, CIDEr, METEOR, SIM
SFHOTEL	raw data	TER, BLEU1-4, ROUGE, LEPOR, CIDEr, METEOR, SIM	METEOR	N/A
SFREST	raw data	SIM	LEPOR	N/A
	quant. data	TER, BLEU1-4, ROUGE, NIST, LEPOR, CIDEr, METEOR SIM	N/A	N/A

Table 4: Metrics predicting relative human rating with significantly higher accuracy than a random baseline.

Discussion: Our data differs from the one used in previous work (Vedantam et al., 2015; Kilickaya et al., 2017), which uses explicit relative rankings ("*Which output do you prefer?*"), whereas we compare two Likert-scale ratings. As such, we have 3 possible outcomes (allowing ties). This way, we can account for equally valid system outputs, which is one of the main drawbacks of forced-choice approaches (Hodosh and Hockenmaier, 2016). Our results are akin to previous work: Kilickaya et al. (2017) report results between 60-74% accuracy for binary classification on machine-machine data, which is comparable to our results for 3-way classification.

Still, we observe a mismatch between the ordinal human ratings and the continuous metrics. For example, humans might rate system A and system B both as a 6, whereas BLEU, for example, might assign 0.98 and 1.0 respectively, meaning that BLEU will declare system B as the winner. In order to account for this mismatch, we quantise our metric data to the same scale as the median scores from our human ratings.[9] Applied to SFREST, where we previously got our worst results, we can see an improvement for predicting *informativeness*, where all WBMs now perform significantly better than the random baseline (see Table 4). In the future, we will investigate related discriminative approaches, e.g. (Hodosh and Hockenmaier, 2016; Kannan and Vinyals, 2017), where the task is simplified to distinguishing correct from incorrect output.

8 Error Analysis

In this section, we attempt to uncover why automatic metrics perform so poorly.

8.1 Scales

We first explore the hypothesis that metrics are good in distinguishing extreme cases, i.e. system outputs which are rated as clearly good or bad by the human judges, but do not perform well for utterances rated in the middle of the Likert scale, as suggested by Kilickaya et al. (2017). We 'bin' our data into three groups: *bad*, which comprises low ratings (≤2); *good*, comprising high ratings (≥5); and finally a group comprising *average* ratings.

We find that utterances with low human ratings of *informativeness* and *naturalness* correlate significantly better ($p < 0.05$) with automatic metrics than those with average and good human ratings. For example, as shown in Figure 3, the correlation between WBMs and human ratings for utterances with low *informativeness* scores ranges between $0.3 \leq \rho \leq 0.5$ (moderate correlation), while the highest correlation for utterances of average and high informativeness barely reaches $\rho \leq 0.2$ (very weak correlation). The same pattern can be observed for correlations with *quality* and *naturalness* ratings.

This discrepancy in correlation results between low and other user ratings, together with the fact that the majority of system outputs are rated "good" for informativeness (79%), naturalness (64%) and quality (58%), whereas low ratings do not exceed 7% in total, could explain why the overall correlations are low (Section 7) despite the observed trends in relationship between average system-level performance scores (Section 6). It also explains why the RNNLG system, which contains very few instances of low user ratings, shows poor correlation between human ratings and automatic metrics.

[9] Note that this mismatch can also be accounted for by continuous rating scales, as suggested by Belz and Kow (2011).

No.	system	MR	system output	human reference	WOMs	SIM	humans
1	LOLS	inform(name = the donatello, hasinternet = yes)	well there is a hotel with internet access called the donatello	the donatello has internet	1.4	5	6
2	LOLS	inform_nomatch(area = embarcadero, kidsallowed= yes, pricerange = expensive)	i but i but i but i but i but i but i but i but i	unfortunately i could not find any expensive restaurants in embarcadero that allow kids.	1.1	1	1
3	TGEN	inform(name = X, area = riverside, eattype = restaurant, food = fastfood, pricerange = cheap)	x is a restaurant on the riverside called located at the riverside and at is	x is a cheap fastfood restaurant located near the riverside	2.4	4	1
4	RNNLG	inform_nomatch(kidsallowed = yes, food = moroccan)	i am sorry, i did not find any restaurants that allows kids and serve moroccan.	sorry, there are no restaurants allowing kids and serving moroccan food	1.85	4	5

Table 5: Example pairs of MRs and system outputs from our data, contrasting the average of word-overlap metrics (normalised in the 1-6 range) and semantic similarity (SIM) with human ratings (median of all measures).

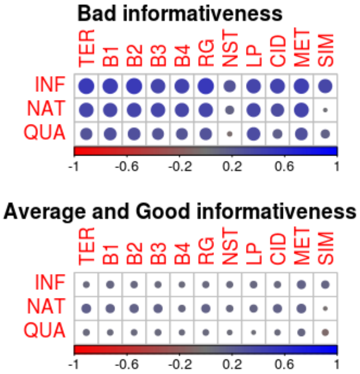

Figure 3: Correlation between automatic metrics (WBMs) and human ratings for utterances of bad informativeness (top), and average and good informativeness (bottom).

8.2 Impact of Target Data

Characteristics of Data: In Section 7.1, we observed that datasets have a significant impact on how well automatic metrics reflect human ratings. A closer inspection shows that BAGEL data differs significantly from SFREST and SFHOTEL, both in terms of grammatical and MR properties. BAGEL has significantly shorter references both in terms of number of characters and words compared to the other two datasets. Although being shorter, the words in BAGEL references are significantly more often polysyllabic. Furthermore, BAGEL only consists of utterances generated from *inform* MRs, while SFREST and SFHOTEL also have less complex MR types, such as *confirm*, *goodbye*, etc. Utterances produced from *inform* MRs are significantly longer and have a significantly higher correlation with human ratings of *informativeness* and *naturalness* than *non-inform* utterance types. In other words, BAGEL is the most complex dataset to generate from. Even though it is more complex, metrics perform most reliably on BAGEL here (note that the correlation is still only weak). One possible explanation is that BAGEL only contains two human references per MR, whereas SFHOTEL and SFREST both contain 5.35 references per MR on average. Having more references means that WBMs naturally will return higher scores ('anything goes'). This problem could possibly be solved by weighting multiple references according to their quality, as suggested by (Galley et al., 2015), or following a reference-less approach (Specia et al., 2010).

Quality of Data: Our corpora contain crowdsourced human references that have grammatical errors, e.g. "*Fifth Floor does not allow childs*" (SFREST reference). Corpus-based methods may pick up these errors, and word-based metrics will rate these system utterances as correct, whereas we can expect human judges to be sensitive to ungrammatical utterances. Note that the parsing score (while being a crude approximation of grammaticality) achieves one of our highest correlation results against human ratings, with $|\rho| = .31$. Grammatical errors raise questions about the quality of the training data, especially when being crowdsourced. For example, Belz and Reiter (2006) find that human experts assign low rankings to their original corpus text. Again, weighting (Galley et al., 2015) or reference-less approaches (Specia et al., 2010) might remedy this issue.

8.3 Example-based Analysis

As shown in previous sections, word-based metrics moderately agree with humans on bad quality output, but cannot distinguish output of good or medium quality. Table 5 provides examples from

Study	Dimension of human ratings Sentence Planning	 Surface Realisation	Domain
this paper	weak positive ($\rho = 0.33$, WPS)	weak negative ($\rho = 0. - 31$, parser)	NLG, restaurant/hotel search
(Reiter and Belz, 2009)	none	strong positive (Pearson's $r = 0.96$, NIST)	NLG, weather forecast
(Stent et al., 2005)	weak positive ($\rho = 0.47$, LSA)	negative ($\rho = -0.56$, NIST)	paraphrasing of news
(Liu et al., 2016)	weak positive ($\rho = 0.35$, BLEU-4)	N/A	dialogue/Twitter pairs
(Elliott and Keller, 2014)	positive ($\rho = 0.53$, METEOR)	N/A	image caption
(Kilickaya et al., 2017)	positive ($\rho = 0.64$, SPICE)	N/A	image caption
(Cahill, 2009)	N/A	negative ($\rho = -0.64$, ROUGE)	NLG, German news texts
(Espinosa et al., 2010)	weak positive ($\rho = 0.43$, TER)	positive ($\rho = 0.62$, BLEU-4)	NLG, news texts

Table 6: Best correlation results achieved by our and previous work. Dimensions targeted towards Sentence Planning include 'accuracy', 'adequacy', 'correctness', 'informativeness'. Dimensions for Surface Realisation include 'clarity', 'fluency', 'naturalness'.

our three systems.[10] Again, we observe different behaviour between WOMs and SIM scores. In Example 1, LOLS generates a grammatically correct English sentence, which represents the meaning of the MR well, and, as a result, this utterance received high human ratings (median = 6) for *informativeness, naturalness* and *quality*. However, WOMs rate this utterance low, i.e. scores of BLEU1-4, NIST, LEPOR, CIDEr, ROUGE and METEOR normalised into the 1-6 range all stay below 1.5. This is because the system-generated utterance has low overlap with the human/corpus references. Note that the SIM score is high (5), as it ignores human references and computes distributional semantic similarity between the MR and the system output. Examples 2 and 3 show outputs which receive low scores from both automatic metrics and humans. WOMs score these system outputs low due to little or no overlap with human references, whereas humans are sensitive to ungrammatical output and missing information (the former is partially captured by GBMs). Examples 2 and 3 also illustrate inconsistencies in human ratings since system output 2 is clearly worse than output 3 and both are rated by human with a median score of 1. Example 4 shows an output of the RNNLG system which is semantically very similar to the reference (SIM=4) and rated high by humans, but WOMs fail to capture this similarity. GBMs show more accurate results for this utterance, with mean of readability scores 4 and parsing score 3.5.

9 Related Work

Table 6 summarises results published by previous studies in related fields which investigate the relation between human scores and automatic metrics. These studies mainly considered WBMs, while we are the first study to consider GBMs. Some studies ask users to provide separate ratings for surface realisation (e.g. asking about 'clarity' or 'fluency'), whereas other studies focus only on sentence planning (e.g. 'accuracy', 'adequacy', or 'correctness'). In general, correlations reported by previous work range from weak to strong. The results confirm that metrics can be reliable indicators at system-level (Reiter and Belz, 2009), while they perform less reliably at sentence-level (Stent et al., 2005). Also, the results show that the metrics capture realization better than sentence planning. There is a general trend showing that best-performing metrics tend to be the more complex ones, combining word-overlap, semantic similarity and term frequency weighting. Note, however, that the majority of previous works do not report whether any of the metric correlations are significantly different from each other.

10 Conclusions

This paper shows that state-of-the-art automatic evaluation metrics for NLG systems do not sufficiently reflect human ratings, which stresses the need for human evaluations. This result is opposed to the current trend of relying on automatic evaluation identified in (Gkatzia and Mahamood, 2015).

A detailed error analysis suggests that automatic metrics are particularly weak in distinguishing outputs of medium and good quality, which can be partially attributed to the fact that human judgements and metrics are given on different scales. We also show that metric performance is data- and system-specific.

Nevertheless, our results also suggest that automatic metrics can be useful for error analysis by helping to find cases where the system is performing poorly. In addition, we find reliable results on

[10] Please note that WBMs tend to match against the reference that is closest to the generated output. Therefore, we only include the closest match in Table 5 for simplicity.

system-level, which suggests that metrics can be useful for system development.

11 Future Directions

Word-based metrics make two strong assumptions: They treat human-generated references as a gold standard, which is *correct* and *complete*. We argue that these assumptions are invalid for corpus-based NLG, especially when using crowdsourced datasets. Grammar-based metrics, on the other hand, do not rely on human-generated references and are not influenced by their quality. However, these metrics can be easily manipulated with grammatically correct and easily readable output that is unrelated to the input. We have experimented with combining WBMs and GBMs using ensemble-based learning. However, while our model achieved high correlation with humans within a single domain, its cross-domain performance is insufficient.

Our paper clearly demonstrates the need for more advanced metrics, as used in related fields, including: assessing output quality within the dialogue context, e.g. (Dušek and Jurčíček, 2016); extrinsic evaluation metrics, such as NLG's contribution to task success, e.g. (Rieser et al., 2014; Gkatzia et al., 2016; Hastie et al., 2016); building discriminative models, e.g. (Hodosh and Hockenmaier, 2016), (Kannan and Vinyals, 2017); or reference-less quality prediction as used in MT, e.g. (Specia et al., 2010). We see our paper as a first step towards reference-less evaluation for NLG by introducing grammar-based metrics. In current work (Dušek et al., 2017), we investigate a reference-less quality estimation approach based on recurrent neural networks, which predicts a quality score for a NLG system output by comparing it to the source meaning representation only.

Finally, note that the datasets considered in this study are fairly small (between 404 and 2.3k human references per domain). To remedy this, systems train on de-lexicalised versions (Wen et al., 2015), which bears the danger of ungrammatical lexicalisation (Sharma et al., 2016) and a possible overlap between testing and training set (Lampouras and Vlachos, 2016). There are ongoing efforts to release larger and more diverse data sets, e.g. (Novikova et al., 2016, 2017).

Acknowledgements

This research received funding from the EPSRC projects DILiGENt (EP/M005429/1) and MaDrIgAL (EP/N017536/1). The Titan Xp used for this research was donated by the NVIDIA Corporation.

References

Anja Belz and Eric Kow. 2011. Discrete vs. continuous rating scales for language evaluation in NLP. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers – Volume 2*. Association for Computational Linguistics, Portland, OR, USA, pages 230–235. http://aclweb.org/anthology/P11-2040.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy, pages 313–320. http://aclweb.org/anthology/E06-1040.

Aoife Cahill. 2009. Correlating human and automatic evaluation of a German surface realiser. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, Suntec, Singapore, pages 97–100. https://aclweb.org/anthology/P09-2025.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy, pages 249–256. http://aclweb.org/anthology/E06-1032.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pages 138–145. http://dl.acm.org/citation.cfm?id=1289189.1289273.

Ondrej Dušek, Jekaterina Novikova, and Verena Rieser. 2017. Referenceless quality estimation for natural language generation. In *Proceedings of the 1st Workshop on Learning to Generate Natural Language*.

Ondřej Dušek and Filip Jurčíček. 2015. Training a natural language generator from unaligned data. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China, pages 451–461. http://aclweb.org/anthology/P15-1044.

Ondřej Dušek and Filip Jurčíček. 2016. A context-aware natural language generator for dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Los Angeles, CA, USA, pages 185–190. arXiv:1608.07076. http://aclweb.org/anthology/W16-3622.

Ondřej Dušek and Filip Jurčíček. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, pages 45–51. arXiv:1606.05491. http://aclweb.org/anthology/P16-2008.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Baltimore, MD, USA, pages 452–457. http://aclweb.org/anthology/P14-2074.

Dominic Espinosa, Rajakrishnan Rajkumar, Michael White, and Shoshana Berleant. 2010. Further meta-evaluation of broad-coverage surface realization. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 564–574. http://aclweb.org/anthology/D10-1055.

Rudolf Franz Flesch. 1979. *How to write plain English: A book for lawyers and consumers*. HarperCollins.

Thomas Francois and Delphine Bernhard, editors. 2014. *Recent Advances in Automatic Readability Assessment and Text Simplification*, volume 165:2 of *International Journal of Applied Linguistics*. John Benjamins. http://doi.org/10.1075/itl.165.2.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pages 445–450. http://aclweb.org/anthology/P15-2073.

Jesús Giménez and Lluís Màrquez. 2008. A smorgasbord of features for automatic MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Columbus, OH, USA, pages 195–198. http://aclweb.org/anthology/W08-0332.

Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. 2016. Natural language generation enhances human decision-making with uncertain information. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany, pages 264–268. arXiv:1606.03254. http://aclweb.org/anthology/P16-2043.

Dimitra Gkatzia and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005–2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*. Association for Computational Linguistics, Brighton, UK, pages 57–60. https://doi.org/10.18653/v1/W15-4708.

Aaron L. F. Han, Derek F. Wong, and Lidia S. Chao. 2012. LEPOR: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of COLING 2012: Posters*. The COLING 2012 Organizing Committee, Mumbai, India, pages 441–450. http://aclweb.org/anthology/C12-2044.

Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. UMBC_EBIQUITY-CORE: Semantic textual similarity systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM)*. Atlanta, Georgia, volume 1, pages 44–52. http://aclweb.org/anthology/S13-1005.

Helen Hastie, Heriberto Cuayahuitl, Nina Dethlefs, Simon Keizer, and Xingkun Liu. 2016. Why bother? Is evaluation of NLG in an end-to-end Spoken Dialogue System worth it? In *Proceedings of the International Workshop on Spoken Dialogue Systems (IWSDS)*. Saariselkä, Finland.

Micah Hodosh and Julia Hockenmaier. 2016. Focused evaluation for image description with binary forced-choice tasks. In *Proceedings of the 5th Workshop on Vision and Language*. Berlin, Germany, pages 19–28. http://aclweb.org/anthology/W16-3203.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H. Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of NAACL-HLT*. Atlanta, GA, USA, pages 1120–1130. http://aclweb.org/anthology/N13-1132.

Min-Yen Kan, Kathleen R. McKeown, and Judith L. Klavans. 2001. Applying natural language generation to indicative summarization. In *Proceedings of the 8th European Workshop on Natural Language Generation*. Association for Computational Linguistics, Toulouse, France, pages 1–9. https://doi.org/10.3115/1117840.1117853.

Anjuli Kannan and Oriol Vinyals. 2017. Adversarial evaluation of dialogue models. *CoRR* abs/1701.08198. https://arxiv.org/abs/1701.08198.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain. arXiv:1612.07600. https://arxiv.org/abs/1612.07600.

Gerasimos Lampouras and Andreas Vlachos. 2016. Imitation learning for language generation from unaligned data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 1101–1112. http://aclweb.org/anthology/C16-1105.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174. https://doi.org/10.2307/2529310.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Prague, Czech Republic, pages 228–231. http://aclweb.org/anthology/W07-0734.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*. Barcelona, Spain, pages 74–81. http://aclweb.org/anthology/W04-1013.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, TX, USA, pages 2122–2132. arXiv:1603.08023. http://aclweb.org/anthology/D16-1230.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Uppsala, Sweden, pages 1552–1561. http://aclweb.org/anthology/P10-1157.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2016. What to talk about and how? Selective generation using LSTMs with coarse-to-fine alignment. In *Proceedings of NAACL-HLT 2016*. San Diego, CA, USA. arXiv:1509.00838. http://aclweb.org/anthology/N16-1086.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, TX, USA, pages 2109–2115. arXiv:1610.02124. http://aclweb.org/anthology/D16-1228.

Jekaterina Novikova, Ondrej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Saarbrücken, Germany. ArXiv:1706.09254. https://arxiv.org/abs/1706.09254.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing NLG data: Pictures elicit better data. In *Proceedings of the 9th International Natural Language Generation Conference*. Edinburgh, UK, pages 265–273. arXiv:1608.00339. http://aclweb.org/anthology/W16-2302.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, PA, USA, pages 311–318. http://aclweb.org/anthology/P02-1040.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics* 35(4):529–558. https://doi.org/10.1162/coli.2009.35.4.35405.

Verena Rieser, Oliver Lemon, and Simon Keizer. 2014. Natural language generation as incremental planning under uncertainty: Adaptive information presentation for statistical dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22(5):979–993. https://doi.org/10.1109/TASL.2014.2315271.

Shikhar Sharma, Jing He, Kaheer Suleman, Hannes Schulz, and Philip Bachman. 2016. Natural language generation in dialogue using lexicalized and delexicalized data. *CoRR* abs/1606.03632. http://arxiv.org/abs/1606.03632.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas*. Cambridge, MA, USA, pages 223–231. http://mt-archive.info/AMTA-2006-Snover.pdf.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation* 24(1):39–50. https://doi.org/10.1007/s10590-010-9077-2.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *Computational Linguistics and Intelligent Text Processing: 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005. Proceedings*. Springer, Berlin/Heidelberg, pages 341–351. https://doi.org/10.1007/978-3-540-30586-6_38.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the 2015 IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA, pages 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Lina Maria Rojas-Barahona, Pei-hao Su, David Vandyke, and Steve J. Young. 2016. Multi-domain neural network language generation for spoken dialogue systems. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, CA, USA, pages 120–129. arXiv:1603.01232. http://aclweb.org/anthology/N16-1015.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, pages 1711–1721. http://aclweb.org/anthology/D15-1199.

Evan James Williams. 1959. *Regression analysis*. John Wiley & Sons, New York, NY, USA.