GRUEN for Evaluating Linguistic Quality of Generated Text

Wanzheng Zhu and **Suma Bhat**
University of Illinois at Urbana-Champaign, USA
wz6@illinois.edu, spbhat2@illinois.edu

Abstract

Automatic evaluation metrics are indispensable for evaluating generated text. To date, these metrics have focused almost exclusively on the content selection aspect of the system output, ignoring the linguistic quality aspect altogether. We bridge this gap by proposing GRUEN for evaluating Grammaticality, non-Redundancy, focUs, structure and coherENce of generated text.[1] GRUEN utilizes a BERT-based model and a class of syntactic, semantic, and contextual features to examine the system output. Unlike most existing evaluation metrics which require human references as an input, GRUEN is reference-less and requires only the system output. Besides, it has the advantage of being unsupervised, deterministic, and adaptable to various tasks. Experiments on seven datasets over four language generation tasks show that the proposed metric correlates highly with human judgments.[2]

1 Introduction

Automatic evaluation metrics for Natural Language Generation (NLG) tasks reduce the need for human evaluations, which can be expensive and time-consuming to collect. Fully automatic metrics allow faster measures of progress when training and testing models, and therefore, accelerate the development of NLG systems (Chaganty et al., 2018; Zhang et al., 2020; Clark et al., 2019).

To date, most automatic metrics have focused on measuring the content selection between the human references and the model output, leaving linguistic quality to be only indirectly captured (*e.g.*, n-gram and longest common subsequence in ROUGE-N and ROUGE-L respectively (Lin and Hovy, 2003; Lin, 2004), and alignment in METEOR (Banerjee and Lavie, 2005)). Even though the need for an explicit measure of linguistic quality has long been pointed out in Dang (2006); Conroy and Dang (2008), this aspect has remained under-explored barring a few studies that focused on measuring the linguistic quality of a generated piece of text (Pitler et al., 2010; Kate et al., 2010; Xenouleas et al., 2019).

Q1: Grammaticality The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (*e.g.*, fragments, missing components) that make the text difficult to read.
Q2: Non-redundancy There should be no unnecessary repetition in the summary.
Q3: Focus The summary should have a focus; sentences should only contain information that is related to the rest of the summary.
Q4: Structure and Coherence The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

Table 1: Dimensions of linguistic quality as proposed in Dang (2006).

In this paper, we bridge this gap by proposing a novel metric for evaluating the *linguistic quality* of system output. Taking into consideration the guidelines put forth for the Document Understanding Conference (DUC) in Table 1, we evaluate: 1) *Grammaticality* by computing the sentence likelihood and the grammatical acceptability with a BERT-based language representation model (Devlin et al., 2019), 2) *Non-redundancy* by identifying repeated components with inter-sentence syntactic features, 3) *Focus* by examining semantic relatedness between adjacent sentences using Word Mover's Distance (WMD) (Kusner et al., 2015), and 4) *Structure and Coherence* by measuring the Sentence-Order Prediction (SOP) loss with A Lite BERT (Lan et al., 2019).

[1] Following BLEU and ROUGE – blue and red in French, we name our evaluation metric GRUEN – that means green in German.

[2] Our metric is available at https://github.com/WanzhengZhu/GRUEN.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 94–108
November 16 - 20, 2020.

Compared with existing metrics, GRUEN is advantageous in that it is:

- *Most correlated* with human judgments: It achieves the highest correlation with human judgments when compared with other metrics of linguistic quality, demonstrated using seven datasets over four NLG tasks.

- *Reference-less*: Most existing evaluation metrics (*e.g.*, ROUGE, METEOR, MoverScore (Zhao et al., 2019)) require human references for comparison. However, it is only logical to assume that the linguistic quality of a system output should be measurable from the output alone. To that end, GRUEN is designed to be reference-less, and requires only the system output as its input.

- *Unsupervised*: Available supervised metrics (*e.g.*, SUM-QE (Xenouleas et al., 2019)) not only require costly human judgments[3] as supervision for each dataset, but also risk poor generalization to new datasets. In addition, they are non-deterministic due to the randomness in the training process. In contrast, GRUEN is unsupervised, free from training and deterministic.

- *General*: Almost all existing metrics for evaluating the linguistic quality are task-specific (*e.g.*, Pitler et al. (2010) and SUM-QE (Xenouleas et al., 2019) are for text summarization), whereas GRUEN is more generally applicable and performs well in various NLG task settings as we demonstrate empirically.

2 Related Work

The growing interest in NLG has given rise to better automatic evaluation metrics to measure the output quality. We first review the widely used metrics for NLG tasks and then discuss available metrics for evaluating linguistic quality.

2.1 NLG Evaluation Metrics

N-gram-based metrics: BLEU (Papineni et al., 2002), ROUGE (Lin and Hovy, 2003; Lin, 2004) and METEOR (Banerjee and Lavie, 2005; Lavie and Denkowski, 2009; Denkowski and Lavie, 2014) are three most commonly used metrics to measure the n-gram lexical overlap between the human references and the system output in various NLG tasks. To tackle their intrinsic shortcomings (*e.g.*, inability to capture lexical similarities), many variations have been proposed such as NIST (Doddington, 2002), ROUGE-WE (Ng and Abrecht, 2015), ROUGE-G (ShafieiBavani et al., 2018) and METEOR++ 2.0 (Guo and Hu, 2019).

Embedding-based metrics: These metrics utilize neural models to learn dense representations of words (Mikolov et al., 2013; Pennington et al., 2014) and sentences (Ng and Abrecht, 2015; Pagliardini et al., 2018; Clark et al., 2019). Then, the embedding distances of the human references and the system output are measured by cosine similarity or Word Movers Distance (WMD) (Kusner et al., 2015). Among them, MoverScore (Zhao et al., 2019), averaging n-gram embeddings with inverse document frequency, shows robust performance on different NLG tasks.

Supervised metrics: More recently, various supervised metrics have been proposed. They are trained to optimize the correlation with human judgments in the training set. BLEND (Ma et al., 2017) uses regression to combine various existing metrics. RUSE (Shimanaka et al., 2018) leverages pre-trained sentence embedding models. SUM-QE (Xenouleas et al., 2019) encodes the system output by a BERT encoder and then adopts a linear regression model. However, all these supervised metrics not only require costly human judgments for each dataset as input, but also have the risk of poor generalization to new datasets and new domains (Chaganty et al., 2018; Zhang et al., 2020). In contrast, unsupervised metrics require no additional human judgments for new datasets or tasks, and can be generally used for various datasets/tasks.

Task-specific metrics: Some metrics are proposed to measure the specific aspects of the tasks. For instance, in text simplification, SARI (Xu et al., 2016) measures the simplicity gain in the output. In text summarization, most metrics are designed to evaluate the content selection, such as Pyramid (Nenkova and Passonneau, 2004), SUPERT (Gao et al., 2020) and Mao et al. (2020). In dialogue systems, diversity and coherence are assessed in Li et al. (2016a,b) and Dziri et al. (2019). However, these proposed metrics are not generally applicable to the evaluation of other aspects or tasks.

[3] We use "human references" to mean the ground truth output for a given task, and "human judgments" as the manual linguistic quality annotation of a system's output.

2.2 Evaluating Linguistic Quality

Existing metrics have focused mostly on evaluating the aspect of content selection in the system output, while ignoring the aspect of linguistic quality. This suggests the long-standing need for automatic measures of linguistic quality of NLG output, despite requests for further studies in this important direction. For instance, the Text Analysis Conference (TAC)[4] and the Document Understanding Conference (DUC)[5] (Dang, 2006) have motivated the need to automatically evaluate the linguistic quality of summarization since 2006. As another example, Conroy and Dang (2008) have highlighted the downsides of ignoring linguistic quality while focusing on summary content during system evaluation. Additionally, the need for linguistic quality evaluation has been underscored in Dorr et al. (2011); Graham et al. (2013); Novikova et al. (2017); Way (2018); Specia and Shah (2018). The uniqueness of our study is that it bridges the need of an automatic evaluation metric of language quality to enable a more holistic evaluation of language generation systems.

Among the few existing metrics of linguistic quality available in prior studies, the early ones Pitler et al. (2010); Kate et al. (2010) rely only on shallow syntactic linguistic features, such as part-of-speech tags, n-grams and named entities. To better represent the generated output, the recent SUM-QE model (Xenouleas et al., 2019) encodes the system output by a BERT encoder and then adopts a linear regression model to predict the linguistic quality. It shows the state-of-the-art results and is most relevant to our work. However, SUM-QE is a supervised metric, which not only requires costly human judgments as input for each dataset, but also has non-deterministic results due to the intrinsic randomness in the training process. Besides, SUM-QE has been shown to work well with the DUC datasets of the summarization task only (Xenouleas et al., 2019), calling into question its effectiveness for other datasets and tasks. GRUEN, as an unsupervised metric, requires no additional human judgments for new datasets and has been shown to be effective on seven datasets over four NLG tasks.

[4] http://tac.nist.gov/

[5] http://duc.nist.gov/

3 Proposed Metric

In this section, we describe the proposed linguistic quality metric in detail. We define the problem as follows: given a system output S with n sentences $[s_1, s_2, ..., s_n]$, where s_i is any one sentence (potentially among many), we aim to output a holistic score, Y_S, of its linguistic quality. We explicitly assess system output for the four aspects in Table 1 – Grammaticality, Non-redundancy, Focus, and Structure and Coherence. We leave Referential Clarity as suggested in Dang (2006) for future work.

Grammaticality: A system output with a high grammaticality score y_g is expected to be readable, fluent and grammatically correct. Most existing works measure the sentence likelihood (or perplexity) with a language model. We, in addition, explicitly capture whether the sentence is grammatically "acceptable" or not.

We measure y_g using two features: sentence likelihood and grammar acceptance. For a system output S, we first use the Punkt sentence tokenizer (Kiss and Strunk, 2006) to extract its component sentences $s_1, s_2, ..., s_n$. Then, for each sentence $s_i = (w_{i,1}, w_{i,2}, ..., w_{i,k})$, a sequence of words $w_{i,j}$, we measure its sentence likelihood score l_i and grammar acceptance score g_i by a BERT model (Devlin et al., 2019).[6] The choice of BERT is to leverage the contextual features and the masked language model (MLM), which can best examine the word choice. However, BERT can not be directly applied to get the likelihood of a sentence, as it is designed to get the probability of a single missing word. Inspired by Wang and Cho (2019); Wang et al. (2019), we estimate l_i by a unigram approximation of the words in the sentence: $l_i = \sum_j \log p(w_{i,j}|w_{i,1}..., w_{i,j-1}, w_{i,j+1}, ..., w_{i,k})$. By such approximation, l_i can be estimated by computing the masked probability of each word. To obtain the grammar acceptance score g_i, we fine-tune the BERT model on the Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2018), a dataset with 10,657 English sentences labeled as grammatical or ungrammatical from linguistics publications. Finally, scores from both models (*i.e.*, l_i and g_i) are linearly combined to examine the grammaticality of the sentence s_i. The final grammaticality score y_g is obtained by averaging scores of all n

[6] We use the "bert-base-cased" model from: http://huggingface.co/transformers/pretrained_models.html.

component sentences: $y_g = \sum_i (l_i + g_i)/n$.

Non-redundancy: As shown in Dang (2006), non-redundancy refers to having no unnecessary repetition, which takes the form of whole sentences or sentence fragments or noun phrases (*e.g.*, "Bill Clinton") when a pronoun ("he") would suffice across sentences. To calculate the non-redundancy score y_r, we capture repeated components by using four *inter-sentence* syntactic features: 1) string length of the longest common substring, 2) word count of longest common words, 3) edit distance, and 4) number of common words. We compute the four features for each pair of component sentences and there are $\binom{n}{2}$ such pairs in total. For each pair of sentences (s_i, s_j), we count the number of times $m_{i,j}$ that these pairs are beyond a non-redundancy penalty threshold. The penalty threshold for each feature are: <80% string length of the shorter sentence, <80% word count of the shorter sentence, >60% string length of the longer sentence, and <80% word count of the shorter sentence, respectively. Finally, we get $y_r = -0.1 * \sum_{i,j} m_{i,j}$. Note that the non-redundancy penalty threshold and penalty weight are learned empirically from a held-out validation set. We discuss the effectiveness of each feature in detail in Appendix B.1.

Focus: Discourse focus has been widely studied and many phenomena show that a focused output should have related semantics between adjacent sentences (Walker, 1998; Knott et al., 2001; Pitler et al., 2010). We compute the focus score y_f by calculating semantic relatedness for each pair of adjacent sentences (s_i, s_{i+1}). Specifically, we calculate the Word Mover Similarity $wms(s_i, s_{i+1})$ (Kusner et al., 2015) for the sentence pair (s_i, s_{i+1}). If the similarity score is less than the similarity threshold 0.05, we will impose a penalty score -0.1 on the focus score y_f. A focused output should expect $y_f = 0$.

Structure and coherence: A well-structured and coherent output should contain well-organized sentences, where the sentence order is natural and easy-to-follow. We compute the inter-sentence coherence score y_c by a self-supervised loss that focuses on modeling inter-sentence coherence, namely Sentence-Order Prediction (SOP) loss. The SOP loss, proposed by Lan et al. (2019), has been shown to be more effective than the Next Sentence Prediction (NSP) loss in the original BERT (Devlin et al., 2019). We formulate the SOP loss calculation as follows. First, for a system output S, we extract all possible consecutive pairs of segments (*i.e.*, $([s_1, ..., s_i], [s_{i+1}, ..., s_n])$, where $i \in [1, 2, ..., n-1]$). Then, we take as positive examples two consecutive segments, and as negative examples the same two consecutive segments but with their order swapped. Finally, the SOP loss is calculated as the average of the logistic loss for all segments,[7] and the coherence score y_c is the additive inverse number of the SOP loss.

Final score: The final linguistic quality score Y_S is a linear combination of the above four scores: $Y_S = y_g + y_r + y_f + y_c$. Note that the final score Y_S is on a scale of 0 to 1, and all the hyper-parameters are learned to maximize the Spearman's correlation with human judgments for the held-out validation set.

4 Empirical Evaluation

In this section, we evaluate the quality of different metrics on four NLG tasks: 1) *abstractive text summarization*, 2) *dialogue system*, 3) *text simplification* and 4) *text compression*.

Evaluating the metrics: We assess the performance of an evaluation metric by analyzing how well it correlates with human judgments. We, following existing literature, report Spearman's correlation ρ, Kendall's correlation τ, and Pearson's correlation r. In addition, to tackle the correlation non-independence issue (two dependent correlations sharing one variable) (Graham and Baldwin, 2014), we report William's significance test (Williams, 1959), which can reveal whether one metric significantly outperforms the other.

Correlation type: Existing automatic metrics tend to correlate poorly with human judgments at the instance-level, although several metrics have been found to have high system-level correlations (Chaganty et al., 2018; Novikova et al., 2017; Liu et al., 2016). Instance-level correlation is critical in the sense that error analysis can be done more constructively and effectively. In our paper, we primarily analyze the instance-level correlations and briefly discuss the system-level correlations.

Baselines: We compare GRUEN with the following baselines:

- **BLEU-best** (Papineni et al., 2002) (best of BLEU-N. It refers to the version that achieves best correlations and is different across datasets.)

[7] We select as the model architecture the pre-trained ALBERT-base model from https://github.com/google-research/ALBERT.

- **ROUGE-best** (Lin, 2004) (best of ROUGE-N, ROUGE-L, ROUGE-W)
- **METEOR** (Lavie and Denkowski, 2009)
- Translation Error Rate (**TER**) (Snover et al., 2006)
- **VecSim** (Pagliardini et al., 2018)
- **WMD-best** (best of Word Mover Distance (Kusner et al., 2015), Sentence Mover Distance (Clark et al., 2019), Sentence+Word Mover Distance (Clark et al., 2019))
- **MoverScore** (Zhao et al., 2019)
- **SUM-QE** (Xenouleas et al., 2019) (we use the "BERT-FT-M-1" model trained on the DUC-2006 (Dang, 2006) and DUC-2007 (Over et al., 2007) datasets)
- **SARI** (Xu et al., 2016) (compared in the text simplification task only)

Note that we do not include Pitler et al. (2010) and Kate et al. (2010), since their metrics rely only on shallow syntactic linguistic features and should probably have no better results than SUM-QE (Xenouleas et al., 2019). Besides, their implementations are not publicly available. For the complete results of BLEU, ROUGE and WMD, please refer to Table 12-15 in Appendix.

4.1 Abstractive Text Summarization

Dataset: We evaluate GRUEN for Text Summarization using two benchmark datasets: the *CNN/Daily Mail* dataset (Hermann et al., 2015; Nallapati et al., 2016) and the *TAC-2011* dataset[8].

The *CNN/Daily Mail* dataset contains online news articles paired with multi-sentence summaries (3.75 sentences or 56 tokens on average). We obtain the human annotated linguistic quality scores from Chaganty et al. (2018) and use the 2,086 system outputs from 4 neural models. Each system output has human judgments on a scale from 1-3 for: *Grammar*, *Non-redundancy* and *Overall* linguistic quality of the summary using the guideline from the DUC summarization challenge (Dang, 2006). In addition, it measures the number of *Post-edits* to improve the summary quality. For all human judgments except *Post-edits*, higher scores indicate better quality.

The *TAC-2011* dataset, from the Text Analysis Conference (TAC), contains 4488 data instances (4.43 sentences or 94 tokens on average). It has 88 document sets and each document set includes 4 human reference summaries and 51 summarizers. We report correlation results on the *Readability* score, which measures the linguistic quality according to the guideline in Dang (2006).

Results: Instance-level correlation scores are summarized in Table 2. As expected, all the baseline approaches except SUM-QE perform poorly because they do not aim to measure linguistic quality explicitly. We note that most of the baselines are highly unstable (and not robust) across the different datasets. For instance, BLEU performs relatively well on TAC-2011 but poor on CNN/Daily Mail, while WMD performs relatively well on CNN/Daily Mail but poor on TAC-2011. GRUEN outperforms SUM-QE on all aspects except the Grammar of CNN/Daily Mail, where they have comparable performance. We performed a set of William's tests for the significance of the differences in performance between GRUEN and SUM-QE for each linguistic score and each correlation type. We found that the differences were significant ($p < 0.01$) in all cases expect the Grammar of CNN/Daily Mail, as shown in Table 8 in Appendix.

4.2 Dialogue System

Dataset: We use three task-oriented dialogue system datasets: BAGEL (Mairesse et al., 2010), SFHOTEL (Wen et al., 2015) and SFREST (Wen et al., 2015), which contains 404, 875 and 1181 instances respectively. Each system output receives *Naturalness* and *Quality* scores (Novikova et al., 2017). *Naturalness* measures how likely a system utterance is generated by native speakers. *Quality* measures how well a system utterance captures fluency and grammar.

Results (Table 3): GRUEN outperforms all other metrics by a significant margin. Interestingly, no metric except GRUEN produces even a moderate correlation with human judgments, regardless of dataset or aspect of human judgments. The finding agrees with the observations in Wen et al. (2015); Novikova et al. (2017); Zhao et al. (2019), where Novikova et al. (2017) attributes the poor correlation to the unbalanced label distribution. Moreover, we analyze the results further in Appendix A.3 in an attempt to interpret them.

4.3 Text Simplification

Dataset: We use a benchmark text simplification dataset with 350 data instances, where each instance has one system output and eight human ref-

[8] http://tac.nist.gov/

	CNN/Daily Mail								TAC-2011	
	Overall		Grammar		Non-redun		Post-edits		Readability	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
BLEU-best	0.17	0.18	0.11	0.12	0.17	0.20	-0.21	-0.29	0.26	0.38
ROUGE-best	0.17	0.19	0.11	0.13	0.20	0.23	-0.24	-0.32	0.25	0.36
METEOR	0.17	0.18	0.10	0.12	0.20	0.22	-0.25	-0.28	0.24	0.32
TER	-0.04	-0.03	0.03	0.02	-0.07	-0.08	0.08	0.08	0.21	0.34
VecSim	0.16	0.19	0.09	0.12	0.18	0.22	-0.24	-0.34	0.16	0.33
WMD-best	0.26	0.24	0.20	0.21	0.26	0.23	-0.29	-0.26	0.15	0.25
MoverScore	0.24	0.26	0.15	0.17	0.28	0.32	-0.32	-0.40	0.29	0.40
SUM-QE	0.46	0.48	0.41	**0.41**	0.45	0.44	-0.51	-0.43	**0.40**	0.41
GRUEN	**0.52**	**0.54**	**0.43**	0.40	**0.52**	**0.58**	**-0.60**	**-0.58**	**0.40**	**0.45**

Table 2: Instance-level Spearman's ρ and Pearson's r correlations on the CNN/Daily Mail and TAC-2011 datasets.

	BAGEL				SFHOTEL				SFREST			
	Naturalness		Quality		Naturalness		Quality		Naturalness		Quality	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
BLEU-best	0.03	0.04	0.02	0.05	0.00	0.07	-0.10	-0.02	0.03	0.03	-0.03	-0.02
ROUGE-best	0.11	0.13	0.10	0.12	-0.02	0.02	-0.12	-0.07	0.02	0.03	-0.06	-0.04
METEOR	0.02	0.03	0.05	0.05	-0.04	0.02	-0.14	-0.07	0.03	0.04	-0.01	0.00
TER	0.11	0.15	0.11	0.15	-0.01	-0.02	-0.05	-0.03	0.01	-0.01	-0.06	-0.08
VecSim	0.03	0.05	0.05	0.07	-0.03	0.04	-0.15	-0.06	0.02	0.02	-0.05	-0.05
WMD-best	0.03	0.05	0.05	0.08	-0.02	0.00	-0.12	-0.07	0.03	0.05	-0.05	0.00
MoverScore	0.07	0.10	0.06	0.10	-0.03	0.02	-0.12	-0.06	0.02	0.02	-0.04	-0.02
SumQE	0.14	0.17	0.13	0.16	0.23	0.30	0.16	0.24	0.09	0.11	0.11	0.13
GRUEN	**0.22**	**0.32**	**0.19**	**0.26**	**0.44**	**0.48**	**0.44**	**0.51**	**0.24**	**0.25**	**0.27**	**0.27**

Table 3: Instance-level Spearman's ρ and Pearson's r correlations on the BAGEL, SFHOTEL and SFREST datasets.

	ρ	τ	r
BLEU-best	0.55	0.40	0.58
ROUGE-best	0.61	0.45	0.64
METEOR	0.63	0.47	**0.67**
TER	0.55	0.40	0.56
VecSim	0.47	0.34	0.53
WMD-best	0.43	0.31	0.33
MoverScore	0.62	0.46	0.65
SumQE	0.62	0.45	0.64
SARI	0.35	0.25	0.40
GRUEN	**0.65**	**0.49**	0.65

Table 4: Instance-level Spearman's ρ, Kendall's τ and Pearson's r correlations with *Grammar* on the text simplification dataset (Xu et al., 2016).

erences (Xu et al., 2016). Each system output instance receives a human-assigned *Grammar* score.

Results: Table 4 presents the results on the dataset of Xu et al. (2016). We note that both GRUEN and METEOR have the best results. The rest of the baseline metrics have satisfactory results too, such as MoverScore and ROUGE. This is unlike the results from the other datasets where most of the baselines correlate poorly with human judgements. A likely explanation is that each data instance from Xu et al. (2016) has eight human references. Having multiple human references capture more allowable variations in language quality and therefore, provide a more comprehensive guideline than a single reference. In Section 5.3, we further analyze this phenomenon and discuss how the number of human references affects the results for each evaluation metric.

	ρ	τ	r
BLEU-best	0.21	0.15	0.21
ROUGE-best	0.41	0.29	0.41
METEOR	0.33	0.23	0.32
TER	0.32	0.23	0.33
VecSim	0.22	0.16	0.23
WMD-best	0.23	0.17	0.25
MoverScore	0.34	0.24	0.34
SumQE	0.38	0.23	0.43
GRUEN	**0.50**	**0.37**	**0.52**

Table 5: Instance-level Spearman's ρ, Kendall's τ and Pearson's r correlations with *Grammar* on the text compression dataset (Toutanova et al., 2016).

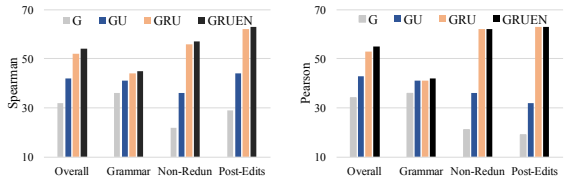

Figure 1: Ablation study on the CNN/Daily Mail Dataset. For better visualization, we present the absolute value of *Post-Edits*.

4.4 Text Compression

Dataset: We use the text compression dataset collected in Toutanova et al. (2016). It has 2955 instances generated by four machine learning systems and each system output instance receives a human-assigned *Grammar* score.

Results (Table 5): We notice that GRUEN outperforms all the other metrics by a significant margin.

5 Discussion

The discussion is primarily conducted for the *text summarization* task considering that GRUEN can measure multiple dimensions in Table 1 of the generated text.

5.1 Ablation study

The results of the ablation analysis (Figure 1) show the effectiveness of G (the Grammaticality module alone), GU (the Grammaticality+focUs modules), GRU (the Grammaticality+non-Redundancy+focUs modules) on the summarization output using the CNN/Daily Mail dataset. We make the following three observations: 1) The *Grammar* score is largely accounted for by our grammaticality module, and only marginally by the others; 2) The focus and non-redundancy module of GRUEN more directly target the *Post-edits* and *Non-redundancy* aspects of linguistic quality; 3) The structure and coherence module does not have significant improvement over the linguistic quality dimensions. One possible reason is that structure and coherence is a high-level feature. It is difficult to be captured by not only the models but also the human annotators. Please refer to Table 6 for an example of a system output with poor structure and coherence.

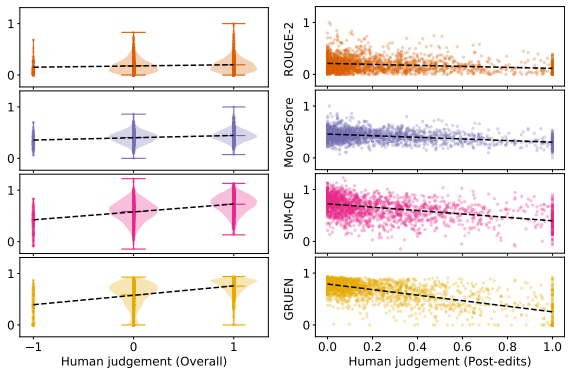

Figure 2: Instance-level distribution of scores for the CNN/Daily Mail dataset. Left shows the *Overall* score distribution on bad (-1), moderate (0) and good (1) outputs. Right shows the scattered *Post-edits* score distribution, which is negatively correlated with the output quality. The dotted line indicates a regression line, which implies the Pearson's correlation r.

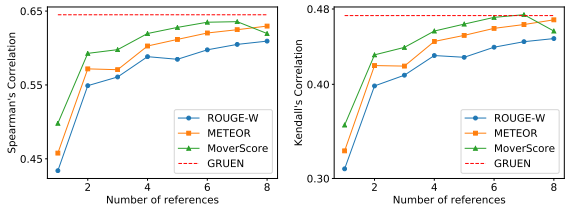

Figure 3: Spearman's Correlation and Kendall's Correlation *v.s.* Number of human references.

5.2 Alignment with Rating Scale

We compared the scores of ROUGE-2, MoverScore, SUM-QE and GRUEN with those of human judgments on outputs of different quality as shown in Figure 2. These are in-line with the findings in Chaganty et al. (2018); Novikova et al. (2017); Zhao et al. (2019) that existing automatic metrics are well correlated with human ratings at the lower end of the rating scale than those in the middle or high end. In contrast, we observe that GRUEN is particularly good at distinguishing high-end cases, *i.e.*, system outputs which are rated as good by the

System output examples	Remarks
(a) Grammaticality: Mr Erik Meldik said the.	Incomplete sentence, and hence has low sentence probability and bad grammar score, captured by the BERT language model.
(b) Grammaticality: Orellana shown red card for throwing grass at Sergio Busquets.	Bad grammar captured by the learned knowledge on the CoLA dataset.
(c) Non-redundancy: The brutal murder of Farkhunda, a young woman in Afghanistan, was burnt and callously chucked into a river in Kabul. The brutal murder of Farkhunda, a young woman in Afghanistan became pallbearers.	Unnecessary repetition (underlined), which can be avoided by using a pronoun (*i.e.*, she). The large overlap between the two sentences is captured by the inter-sentence syntactic features.
(d) Focus: The FDA's Nonprescription Drugs Advisory Committee will meet Oct. Infant cough-and-cold products were approved decades ago without adequate testing in children because experts assumed that children were simply small adults, and that drugs approved for adults must also work in children. Ian Paul, an assistant professor of pediatrics at Penn State College of Medicine who has studied the medicines.	Component sentences are scattered, of different themes or even irrelevant to each other. The sentence embedding similarity of each pair of adjacent sentences is low and thus, results in low Focus score.
(e) Structure and Coherence: Firefighters worked with police and ambulance staff to free the boy, whose leg was trapped for more than half an hour down the hole. It is believed the rubber drain cover had been kicked out of position and within hours, the accident occurred. A 12-year-old schoolboy needed to be rescued after falling down an eight-foot drain in Peterborough.	The output is only a heap of related information, where the component sentences are in a unorganized, wrong or incomprehensible order. Its sentence structure and readability can be much improved if the three component sentences are in the order of 3,1,2.

Table 6: Case study: linguistic quality analysis

	ρ	τ	r
BLEU-best	0.51	0.38	0.61
ROUGE-best	0.52	0.38	0.71
METEOR	0.45	0.30	0.73
TER	0.64	0.46	0.71
VecSim	0.38	0.27	0.62
WMD-best	0.31	0.23	0.60
MoverScore	0.42	0.30	0.66
SUM-QE	0.76	0.63	0.69
GRUEN	**0.87**	**0.69**	**0.85**

Table 7: System-level Spearman's ρ, Kendall's τ and Pearson's r correlations with *Readability* on the TAC-2011 dataset.

human judges.

5.3 Impact of Number of References

Figure 3 shows how the Spearman's correlation of each metric varies with different numbers of human references in the text simplification dataset. It is clear that existing reference-based metrics show better performance with more human references. One possible reason is that the system outputs are compared with more allowable grammatical and semantic variations. These allowable variations could potentially make the reference-based metrics better at distinguishing high-end cases, alleviate the shortcoming in Section 5.2, and thus allow the metrics to perform well. However, in most cases, it is expensive to collect multiple human references for each instance.

5.4 Case Study

Table 6 presents a case study on examples with poor Grammaticality, Non-redundancy, Focus, and Structure and Coherence. In Table 10-11 in the Appendix, we further analyze how non-redundancy is captured by each of the inter-sentence syntactic features, and also present a comparative study for each linguistic dimension.

5.5 System-level Correlation

Our results have shown that GRUEN improves the instance-level correlation performance from poor to moderate. At the system-level too, we observe significant improvements in correlation. Table 7 shows the system-level linguistic quality correlation scores for *Readability* on the TAC-2011 dataset, which consists of 51 systems (*i.e.*, summarizers). At the system level, most baseline metrics have moderate correlations, which aligns with the findings in Chaganty et al. (2018), while GRUEN achieves a high correlation. Note that we do not further study the system-level correlations on other datasets, since they have no more than four systems and thus the correlations are not meaningful to be compared with.

5.6 Limitations and Future Work

GRUEN evaluates non-redundancy by looking for lexical overlap across sentences. However, they still remain unexamined for semantically relevant components that are in different surface forms. Besides, it does not handle intra-sentence redundancy, such as "In 2012, Spain won the European Championships for a second time in 2012.". Another challenging problem is to evaluate the referential clarity as proposed in Dang (2006), which is particularly important for long sentences and multi-sentence outputs. Future work should aim for a more comprehensive evaluation of redundancy and tackle the referential clarity challenge.

6 Conclusion

We proposed GRUEN to evaluate Grammaticality, non-Redundancy, focUs, structure and coherENce of generated text. Without requiring human references, GRUEN achieves the new state-of-the-art results on seven datasets over four NLG tasks. Besides, as an unsupervised metric, GRUEN is deterministic, free from obtaining costly human judgments, and adaptable to various NLG tasks.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *The ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Association for Computational Linguistics (ACL)*.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence movers similarity: Automatic evaluation for multi-sentence texts. In *Association for Computational Linguistics (ACL)*.

John M Conroy and Hoa Trang Dang. 2008. Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In *Association for Computational Linguistics (ACL)*.

Hoa Trang Dang. 2006. Overview of duc 2006. In *Document Understanding Conference (DUC)*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Workshop on Statistical Machine Translation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Association for Computational Linguistics (NAACL)*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *International Conference on Human Language Technology Research*.

Bonnie Dorr, Joseph Olive, John McCary, and Caitlin Christianson. 2011. Machine translation evaluation and optimization. In *Handbook of Natural Language Processing and Machine Translation*, pages 745–843. Springer.

Nouha Dziri, Ehsan Kamalloo, Kory W Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. *arXiv preprint arXiv:1904.03371*.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Association for Computational Linguistics (ACL)*.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *The Seventh Linguistic Annotation Workshop and Interoperability with Discourse*.

Yinuo Guo and Junfeng Hu. 2019. Meteor++ 2.0: Adopt syntactic level paraphrase knowledge into machine translation evaluation. In *The Fourth Conference on Machine Translation*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Neural Information Processing Systems (NIPS)*.

Rohit J Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *International Conference on Computational Linguistics (COLING)*.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Alistair Knott, Jon Oberlander, Mick ODonnell, and Chris Mellish. 2001. Beyond elaboration: The interaction of relations and focus in coherent text. *Text Representation: Linguistic and Psycholinguistic Aspects*, pages 181–196.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning (ICML)*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *North American Association for Computational Linguistics (NAACL)*.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *North American Association for Computational Linguistics (NAACL)*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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 *Empirical Methods in Natural Language Processing (EMNLP)*.

Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. 2017. Blend: a novel combined mt metric based on direct assessmentcasict-dcu submission to wmt17 metrics task. In *The Second Conference on Machine Translation*.

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 *Association for Computational Linguistics (ACL)*.

Yuning Mao, Liyuan Liu, Qi Zhu, Xiang Ren, and Jiawei Han. 2020. Facet-aware evaluation for extractive summarization. In *Association for Computational Linguistics (ACL)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems (NIPS)*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Ça glar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *Computational Natural Language Learning (CoNLL)*.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *North American Association for Computational Linguistics (NAACL)*.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. *Information Processing & Management*, 43(6):1506–1520.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *North American Association for Computational Linguistics (NAACL)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics (ACL)*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Association for Computational Linguistics (ACL)*.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2018. A graph-theoretic summary evaluation for rouge. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In *The Third Conference on Machine Translation: Shared Task Papers*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Association for Machine Translation in the Americas*.

Lucia Specia and Kashif Shah. 2018. Machine translation quality estimation: Applications and future perspectives. In *Translation Quality Assessment*, pages 201–235. Springer.

Kristina Toutanova, Chris Brockett, Ke M Tran, and Saleema Amershi. 2016. A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Joshi Prince Walker. 1998. *Centering theory in discourse*. Oxford University Press.

Alex Wang and Kyunghyun Cho. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model. In *The Workshop on Methods for Optimizing and Evaluating Neural Language Generation*.

Chenguang Wang, Mu Li, and Alexander J Smola. 2019. Language models with transformers. *arXiv preprint arXiv:1904.09408*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Andy Way. 2018. Quality expectations of machine translation. In *Translation Quality Assessment*, pages 159–178. Springer.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Evan James Williams. 1959. *Regression Analysis*, volume 14. Wiley.

Stratos Xenouleas, Prodromos Malakasiotis, Marianna Apidianaki, and Ion Androutsopoulos. 2019. Sum-qe: a bert-based summary quality estimation model. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics (TACL)*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

		ρ	τ	r
CNN/ Daily Mail	**Overall**	***	***	***
	Grammar	*	—	—
	Non-Redun	***	***	***
	Post-edits	***	***	***
TAC-2011	**Readability**	*	**	**
BAGEL	**Naturalness**	0.01	0.07	**
	Quality	0.06	0.17	*
SFHOTEL	**Naturalness**	***	***	***
	Quality	***	***	***
SFREST	**Naturalness**	***	***	***
	Quality	***	***	***
Xu et al. (2016)	**Grammar**	0.33	0.46	—
Toutanova et al. (2016)	**Grammar**	***	***	***

Table 8: William Significance Test on GRUEN against the best baselines for each correlation type and each dataset. *, **, *** indicate the significance level of <0.01, <0.001 and <0.0001 respectively. — indicates GRUEN does not outperform the best baseline.

A Quantitative Analysis

A.1 William's Significance Test

In Table 8, we perform William's significance tests on GRUEN against the best baselines for each linguistic score and each correlation measurement (*e.g.*, SUM-QE for ρ on the *Overall* score of the CNN/Daily Mail dataset, METEOR for r on the *Grammar* score of the dataset in Xu et al. (2016)). We found that the differences are significant ($p < 0.0001$) in 24 out of 39 cases.

A.2 Performance on Reliable Instances

In the human annotation process, each instance receives a score that is the aggregate of multiple people's ratings. Given the subjective nature of the task of annotating for linguistic quality, there are some instances where annotators disagree. To analyze how we perform on reliably coded instances, we show in Table 9 the correlation scores on the instances where all annotators agreed perfectly on the *Overall* score for the CNN/Daily Mail dataset ($N = 1323$). We observe that GRUEN consistently outperforms the baselines on the reliable data instances. Importantly, GRUEN and SUM-QE are better correlated with human judgements on the reliable data instances than on all the data instances.

	Overall			Grammar			Non-redun			Post-edits		
	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ	r
BLEU-best	0.17	0.14	0.20	0.12	0.09	0.15	0.20	0.15	0.23	-0.23	-0.16	-0.33
ROUGE-best	0.17	0.14	0.19	0.13	0.09	0.15	0.20	0.15	0.23	-0.26	-0.18	-0.34
METEOR	0.17	0.13	0.17	0.11	0.09	0.13	0.20	0.15	0.21	-0.26	-0.18	-0.31
TER	-0.01	-0.00	0.01	0.03	0.02	0.04	-0.05	-0.04	-0.04	0.06	0.04	0.07
VecSim	0.17	0.13	0.21	0.11	0.08	0.15	0.20	0.15	0.26	-0.27	-0.18	-0.40
WMD-best	0.27	0.21	0.26	0.24	0.18	0.25	0.27	0.20	0.25	-0.31	-0.21	-0.29
MoverScore	0.23	0.18	0.25	0.17	0.13	0.20	0.28	0.21	0.32	-0.33	-0.23	-0.41
SUM-QE	0.53	0.43	0.54	0.49	**0.38**	**0.49**	0.47	0.36	0.46	-0.54	-0.38	-0.45
GRUEN	**0.58**	**0.47**	**0.58**	**0.50**	0.37	0.48	**0.62**	**0.48**	**0.66**	**-0.68**	**-0.50**	**-0.64**

Table 9: Instance-level Spearman's ρ, Kendall's τ and Pearson's r correlations on the **reliable** data instances of the CNN/Daily Mail dataset.

Example Outputs	**Feature**
(1): The monkey took a bunch of bananas on the desk. It took a bunch of bananas on the desk.	ABCD
(2): The monkey took a bunch of bananas on the desk. The monkey took a bunch of bananas on the desk, and they are the fruits reserved for the special guests invited tonight.	ABD
(3): The monkey took a bunch of bananas on the desk. The monkey took a large bunch of bananas on the red desk.	CD
(4): The monkey took a bunch of bananas on the desk. It took bunches of banana on the desks.	C

Table 10: Example with poor non-redundancy. The features that contribute to the non-redundancy penalty are labeled on the right.

A.3 Analysis on the Dialogue System Datasets

Table 3 has shown an extremely poor correlation with human ratings for the baseline metrics on the BAGEL, SFHOTEL and SFREST datasets. Novikova et al. (2017) hypothesizes the reason to be the unbalanced label distribution. It turns out that the majority of system outputs are good for *Naturalness* with 64% and *Quality* (58%), whereas bad examples are only 7% in total.[9] Our discussion in Section 5.2 further explains the reason. Existing metrics are bad at assigning high scores to good outputs and thus, have a very poor correlation in such datasets with mostly good examples. In contrast, GRUEN is capable of assigning high scores to good outputs and thus, achieves decent correlation results.

While our correlation results may appear to be slightly different from Table 3 in Novikova et al. (2017), they are in fact the same. The only difference is the result presentation format. Novikova et al. (2017) presents only the best correlation results for each dataset (*i.e.*, BAGEL, SFHOTEL and SFREST) and each NLG system (*i.e.*, TGEN, LOLS and RNNLG), while we present the average correlation score for each dataset. Therefore, in Table 3 of Novikova et al. (2017), a correlation metric that performs well on one NLG system does not mean it performs equally well on another NLG system. As an example of measuring *Informativeness*, BLEU-1 performs well on the TGEN system for the BAGEL dataset, while it performs poorly on the LOLS system for the BAGEL dataset. Therefore, BLEU-1 has only a mediocre correlation score over informativeness for the BAGEL dataset, as presented in our result. The analysis in Novikova et al. (2017) is more focused in that it analyzes different metrics in a more restricted manner, whereas our analysis of metrics is more general in that we compare correlation scores regardless of which NLG system the output was generated from.

[9] In a 6-point scale, *bad* comprises low ratings ($\leq$2), while *good* comprises high ratings ($\geq$5).

B Qualitative Analysis

B.1 Analysis on Non-redundancy

To evaluate the non-redundancy score y_r of a system output, we capture repeated components of a

Example Outputs	Module Scores
(a) Grammaticality: Orellana shown red card for throwing grass at Sergio Busquets.	$y_g = 0.2$
(b) Grammaticality: Orellana was shown a red card for throwing grass at Sergio Busquets.	$y_g = 0.7$
(c) Non-redundancy: The brutal murder of Farkhunda, a young woman in Afghanistan, whose body was burnt and callously chucked into a river in Kabul. The brutal murder of Farkhunda, a young woman in Afghanistan became pallbearers, hoisting the victim's coffin on their shoulders draped with headscarves.	$y_r = -0.4$
(d) Non-redundancy: The brutal murder of Farkhunda, a young woman in Afghanistan, whose body was burnt and callously chucked into a river in Kabul. She became pallbearers, hoisting the victim's coffin on their shoulders draped with headscarves.	$y_r = 0.0$
(e) Focus: The FDA's Nonprescription Drugs Advisory Committee will meet Oct. Infant cough-and-cold products were approved decades ago without adequate testing in children because experts assumed that children were simply small adults, and that drugs approved for adults must also work in children. Ian M. Paul, an assistant professor of pediatrics at Penn State College of Medicine who has studied the medicines.	$y_f = -0.1$
(f) Focus: On March 1, 2007, the Food/Drug Administration (FDA) started a broad safety review of children's cough/cold remedies. They are particularly concerned about use of these drugs by infants. By September 28th, the 356-page FDA review urged an outright ban on all such medicines for children under six. Dr. Charles Ganley, a top FDA official said "We have no data on these agents of what's a safe and effective dose in Children." The review also stated that between 1969 and 2006, 123 children died from taking decongestants and antihistamines. On October 11th, all such infant products were pulled from the markets.	$y_f = 0.0$
(g) Coherence and Structure: Firefighters worked with police and ambulance staff to free the boy, whose leg was trapped for more than half an hour down the hole. It is believed the rubber drain cover had been kicked out of position and within hours, the accident occurred. A 12-year-old schoolboy needed to be rescued after falling down an eight-foot drain in Peterborough.	$y_c = -0.1$
(h) Coherence and Structure: A 12-year-old schoolboy needed to be rescued after falling down an eight-foot drain in Peterborough. Firefighters worked with police and ambulance staff to free the boy, whose leg was trapped for more than half an hour down the hole. It is believed the rubber drain cover had been kicked out of position and within hours, the accident occurred.	$y_c = 0.0$
(i) Overall: The monkey took a bottle of a water bottle in a bid to cool it down with bottle in hand. The monkey is the bottle to its hands before attempting to quench its thirst. It is the the bottle of the bottle in its mouth and a bottle. It's the bottle. A bottle in the water bottle.	$Y_S = 0.0$
(j) Overall: The footage was captured on a warm day in Bali, Indonesia. Tour guide cools monkey down by spraying it with water. Monkey then picks up bottle and casually unscrews the lid. Primate has drink and remarkably spills very little liquid.	$Y_S = 0.8$

Table 11: A comparative study on good and bad example outputs for each linguistic aspect.

pair of sentences by four empirical *inter-sentence* syntactic features: (A) length of longest common substring, (B) length of longest common words, (C) edit distance, and (D) number of common words. Features (A) and (B) focus on continuous word overlap of a pair of sentences. Intuitively, when most characters of a sentence already appears in the other sentence, the system output should probably have a poor non-redundancy score. However, features (A) and (B) fail to make a quality evaluation when the repeated components are of a inflected form (*e.g.*, stemming, lemmatization) or not continuous. To account for the above limitation, we introduce features (C) and (D) that measures the edit distance and the number of common words respectively.

To gain more intuition, we present a few examples of poor non-redundancy in Table 10. The features that contribute to the non-redundancy penalty are labeled on the right. Case (1) has two almost identical sentences and therefore, captured by all four features. However, when the word lengths of the two sentences differ a lot, feature (C) is no longer effective as shown in case (2). In case (3) where the word overlap is not continuous (*i.e.*, "The monkey took a" and "bunch of bananas on the"), the non-redundancy can only be detected by features (C) and (D). In case (4), the composing words are of an inflected form and thus, can not be captured by exact word matching features (*i.e.*, features (A), (B), (D)). As such, we have the four features to complement each other and aim to capture non-redundancy well.

B.2 Comparative Study

Table 11 presents a comparative study on good and bad examples for each linguistic quality aspect, together with their corresponding module scores. Besides, we compare two examples with good and bad overall linguistic quality scores.

C Complete Results

We present the complete results of BLEU, ROUGE and WMD for all tasks in Table 12-15.

	ρ	τ	r
BLEU-1	0.38	0.28	0.41
BLEU-2	0.47	0.33	0.49
BLEU-3	0.52	0.37	0.55
BLEU-4	0.55	0.40	0.58
ROUGE-1	0.51	0.37	0.56
ROUGE-2	0.54	0.39	0.58
ROUGE-3	0.52	0.38	0.55
ROUGE-4	0.50	0.36	0.51
ROUGE-L	0.56	0.40	0.59
ROUGE-W	0.61	0.45	0.64
WMD	0.43	0.31	0.33
SMD	0.30	0.21	0.30
S+WMD	0.40	0.29	0.34

Table 12: Instance-level Spearman's ρ, Kendall's τ and Pearson's r correlations with *Grammar* on the text simplification dataset (Xu et al., 2016).

	ρ	τ	r
BLEU-1	0.07	0.05	0.17
BLEU-2	0.12	0.08	0.18
BLEU-3	0.17	0.12	0.19
BLEU-4	0.21	0.15	0.21
ROUGE-1	0.21	0.15	0.24
ROUGE-2	0.33	0.24	0.34
ROUGE-3	0.35	0.26	0.37
ROUGE-4	0.35	0.25	0.36
ROUGE-L	0.39	0.28	0.37
ROUGE-W	0.41	0.29	0.41
WMD	0.18	0.13	0.16
SMD	0.23	0.17	0.25
S+WMD	0.20	0.14	0.21

Table 13: Instance-level Spearman's ρ, Kendall's τ and Pearson's r correlations with *Grammar* on the text compression dataset (Toutanova et al., 2016).

	CNN/Daily Mail								TAC-2011	
	Overall		Grammar		Non-redun		Post-edits		Readability	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
BLEU-1	0.07	0.08	0.05	0.05	0.06	0.06	-0.08	-0.10	0.17	0.34
BLEU-2	0.13	0.14	0.09	0.09	0.13	0.14	-0.16	-0.20	0.21	0.35
BLEU-3	0.16	0.18	0.10	0.12	0.17	0.19	-0.21	-0.27	0.24	0.36
BLEU-4	0.17	0.18	0.11	0.12	0.17	0.20	-0.21	-0.29	0.26	0.38
ROUGE-1	0.17	0.19	0.11	0.13	0.20	0.23	-0.24	-0.32	0.25	0.36
ROUGE-2	0.14	0.13	0.09	0.10	0.15	0.15	-0.18	-0.21	0.25	0.26
ROUGE-3	0.12	0.10	0.08	0.09	0.13	0.11	-0.16	-0.16	0.24	0.19
ROUGE-4	0.10	0.08	0.08	0.08	0.11	0.09	-0.14	-0.13	0.20	0.15
ROUGE-L	0.12	0.13	0.10	0.12	0.11	0.12	-0.17	-0.19	0.25	0.36
ROUGE-W	0.14	0.14	0.10	0.12	0.13	0.14	-0.18	-0.19	0.26	0.34
WMD	0.18	0.11	0.12	0.10	0.19	0.11	-0.23	-0.15	0.19	0.17
SMD	0.26	0.24	0.20	0.21	0.26	0.23	-0.29	-0.26	0.15	0.25
S+WMD	0.21	0.17	0.15	0.15	0.22	0.17	-0.26	-0.21	0.19	0.24

Table 14: Instance-level Spearman's ρ and Pearson's r correlations on the CNN/Daily Mail and TAC-2011 datasets.

	BAGEL				SFHOTEL				SFREST			
	Naturalness		Quality		Naturalness		Quality		Naturalness		Quality	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
BLEU-1	-0.02	-0.02	-0.02	-0.01	0.03	0.11	-0.04	0.04	0.03	0.03	-0.03	-0.02
BLEU-2	0.00	0.00	-0.01	0.01	0.01	0.09	-0.08	0.00	0.03	0.02	-0.03	-0.03
BLEU-3	0.01	0.03	0.01	0.03	0.00	0.08	-0.10	-0.01	0.03	0.02	-0.03	-0.03
BLEU-4	0.03	0.04	0.02	0.05	0.00	0.07	-0.10	-0.02	0.03	0.03	-0.03	-0.02
ROUGE-1	0.10	0.12	0.10	0.12	-0.01	0.06	-0.11	-0.03	0.02	0.01	-0.05	-0.04
ROUGE-2	0.11	0.13	0.10	0.12	-0.02	0.02	-0.12	-0.07	0.02	0.03	-0.06	-0.04
ROUGE-3	0.08	0.10	0.07	0.09	-0.03	0.01	-0.12	-0.06	0.01	0.04	-0.06	-0.03
ROUGE-4	0.04	0.09	0.04	0.08	-0.04	0.00	-0.12	-0.06	0.02	0.05	-0.04	-0.01
ROUGE-L	0.08	0.10	0.09	0.11	-0.01	0.07	-0.11	-0.03	0.01	0.01	-0.06	-0.04
ROUGE-W	0.08	0.10	0.08	0.10	-0.02	0.04	-0.12	-0.05	0.05	0.05	-0.03	-0.02
WMD	0.03	0.05	0.05	0.08	-0.02	0.00	-0.12	-0.07	0.03	0.05	-0.05	0.00
SMD	0.00	0.04	0.02	0.07	0.00	0.01	-0.09	-0.06	-0.01	0.03	-0.07	-0.01
S+WMD	0.02	0.05	0.04	0.08	-0.01	0.00	-0.11	-0.07	0.02	0.05	-0.06	-0.01

Table 15: Instance-level Spearman's ρ and Pearson's r correlations on the BAGEL, SFHOTEL and SFREST datasets.