GRUENer Fields of NLG: Assessing Novel Auto-Eval Metric for GPT-2 Generated Text

Manpreet Khural
UC Berkeley School of Information
manpreet.khural@berkeley.edu

Abstract

As Natural Language Generation continues its upwards development trajectory, we begin to interact with increasing volumes of synthetically generated text. From improved chat bots to fun experiments in style transfer,[1] evaluating the "goodness" of outputs of such models proves holistically elusive. The gold standard for measuring the quality of the generated text remains centrally based on human judgment, requiring higher levels of funding as well as delays to accumulate feedback. Automatic metrics such as BLEU and ROUGE have proven effective for specific narrow tasks such as sentence-based machine translation[2] , requiring no training to assess the quality of text. On this front, a newly metric, GRUEN, keeping with the color scheme has been developed to better capture dimensions of linguistic quality. An evaluation of this metric on real web text compared to GPT-2 generated text shows that this proposed metric is not reliable enough for widespread use, though the framework has potential.

Keywords*:* NLG, Evaluation Metrics, Unsupervised

Introduction

Performing intrinsic evaluations of generated text is notoriously difficult as compared to extrinsic. The former seeks to understand the inherent quality of text either in terms of components such as fluency and coherence. The latter can be captured by the implementation of the text output in a task that has a measurable result. There is no sole agreed objective criterion for goodness of text. Generally, these evaluation approaches can be broken down into the following: (1) Human-centric evaluation, (2) Supervised Metrics that use smaller sets of human reviews to build semantic similarity, factual correctness, or language model based systems, and (3) Unsupervised Automatic Evaluation Metrics.

Currently different metrics target specific tasks largely rely on the task the NLG system is attempting to model. Human evaluation remains the one constant, but even that requires careful

[1] Qian, Y., 2020. Story-level Text Style Transfer: A Proposal
[2] Asli Celikyilmaz, Elizabeth Clark and Jianfeng Gao. 2021. Evaluation of Text Generation: A Survey.

assessment methodology to have reliability. The Turing Test is one such methodology often used, measuring the overall “distance” between perceived quality of text. These pose a great deal of expense and consume a lot of time for researchers looking to evaluate their models. As such, the Automatic Evaluation Metrics are far more intriguing. Recently a new metric called GRUEN has been proposed that seeks to better capture linguistic quality and outperform other similar models across many more tasks. This deserves some further evaluation and assessment.

Background

The novelty in the GRUEN model lies in the fact that it seeks to marry many different dimensions of linguistic quality. Other research also explores creating evaluation metrics for very specific linguistic features. Seeking to generate their own rap lyrics the DopeLearning team, break down the linguistic and qualitative features of Rhyming and create a rhyme density measure to capture it (validated by human review) as well as Song Structure, understanding that verses and choruses are distinctly different and consist of predictable syllabic length[3].

The GRUEN metric uses: (1) Grammaticality or how strong the grammar is, (2) Non-redundancy or avoiding repetition, (3) Focus or information relatedness, and (4) Structure and Coherence or flow[4]. Metrics already exist that measure some of this such as fluency and semantic similarity, but this method seeks to marry all of these factors together for a final score: $Y_S = y_g + y_r + y_f + y_c$[5] bound between 0 and 1.

Methods

GRUEN uses mixed approaches such as transformers for grammar and semantic similarity. It claims strong scores in abstractive text summarization, dialogue system, text simplification, and text compression, evaluated using datasets with human reviews and the correlation the score had to these. This research does not have access to human reviews so it is estimated. Using an OpenAI GPT-2 generated text dataset that uses WebText data to train on, we can estimate human scoring by comparing the real, web scraped text to the generated tests. Using the smallest GPT-2 model, we assume the text would score poorly by humans where the real text would score significantly higher if the metric is valid.

We run GRUEN on these two datasets with a modification to output the individual linguistic qualities being captured. We seek to determine how the total metric score and component scores relate to the proxy human scores the differing datasets provide. We examine the comparative

[3] Eric Malmi, Pyry Takala, Hannu Toivonen, Tapani Raiko and Aristides Gionis. 2016. DopeLearning.

[4] Paul Over, Hoa Dang, and Donna Harman. 2007. Duc
in context.

[5]

distributions of each score to determine whether an optimization could be made. We also run regressions to examine the explicit relationship between the scores and being able to discriminate between which text is real or generated. We finish by examining specific outliers. We do not implement a model change but assess where and it could be done.

Results

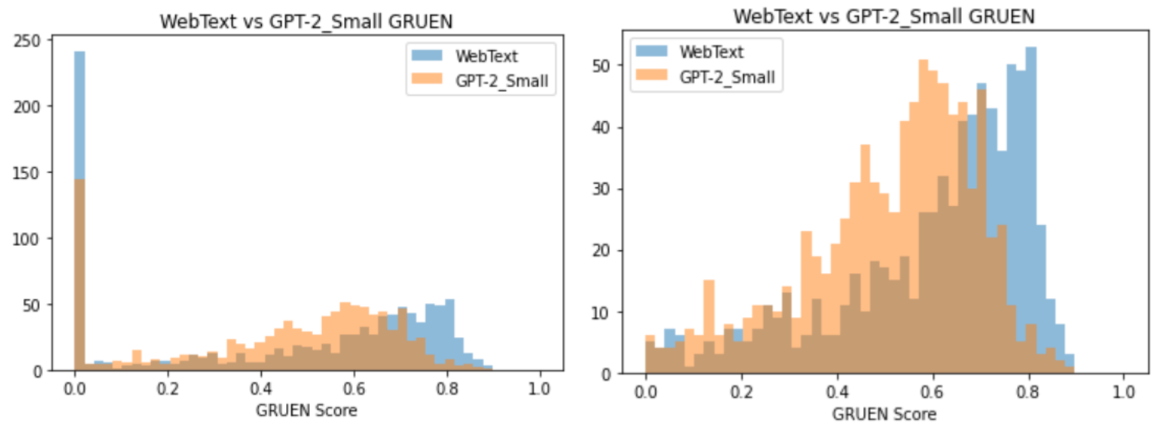


Figure 1. Due to strongly negative redundancy scores and an aggregation function that lower-bounds the score at 0, we notice many 0 values.

The comparison between GRUEN scores for the real and generated highlight a potential issue with the aggregation method. 23.6% of all real text scores and 13.8% of generated text scores were zeros. This contradicts the relationship in the other data points that indicate as expected that WebText would score better. The metric holds up in that regard, but the score aggregation function seems dubious.

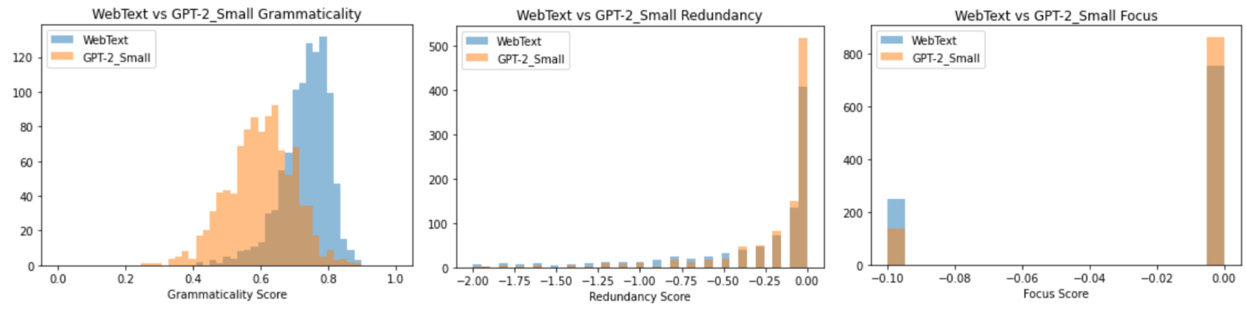
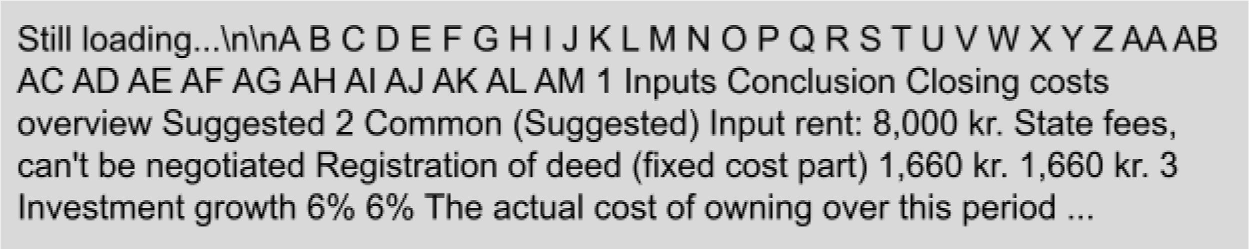


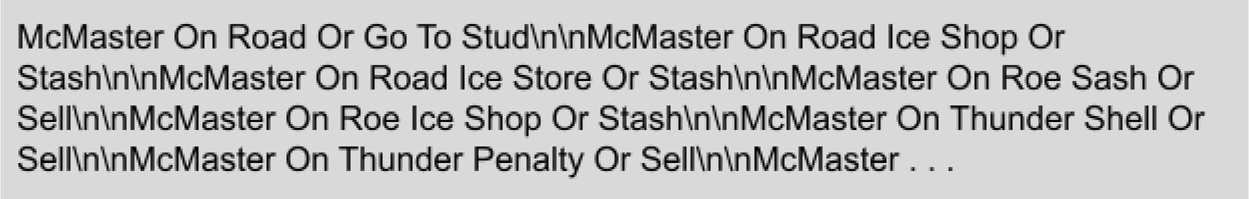
Figure 2. Grammaticality, which already has incorporated Structure and Coherence shows stronger potential performance. The redundancy score and focus score seem to have no valuable correlation with the relative text quality we have estimated for each dataset.

Examining the other scores, Grammaticality performs well, using a BERT Transformer backbone that comes pre-trained with finetuned weights specifically for this linguistic feature. It mirrors the overall score. Scores of Focus and Redundancy, in particular, seem to not indicate that GPT-2 output suffers on these qualities. The measurement for these is particularly poor. Focus only either has values of -0.1 or 1.0. Redundancy compromises the data heavily as it has huge outliers. WebText has one with a score of -419.7 and GPT-2 has one with a massive -3138.8. To be a stronger option, GRUEN would have to be more robust to these.



Still loading...\n\nA B C D E F G H I J K L M N O P Q R S T U V W X Y Z AA AB AC AD AE AF AG AH AI AJ AK AL AM 1 Inputs Conclusion Closing costs overview Suggested 2 Common (Suggested) Input rent: 8,000 kr. State fees, can't be negotiated Registration of deed (fixed cost part) 1,660 kr. 1,660 kr. 3 Investment growth 6% 6% The actual cost of owning over this period ...

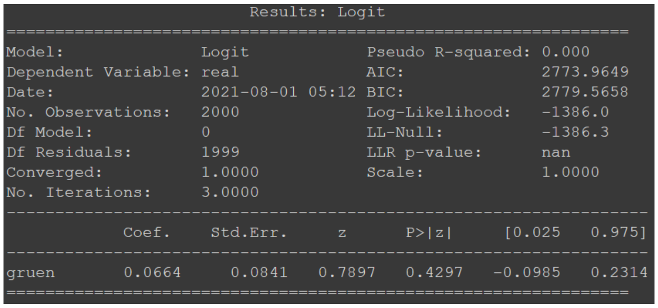
Figure 3. WebText Redundancy Score minimum outlier at -419.7



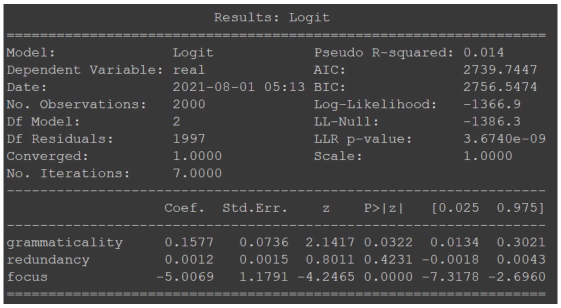
McMaster On Road Or Go To Stud\n\nMcMaster On Road Ice Shop Or Stash\n\nMcMaster On Road Ice Store Or Stash\n\nMcMaster On Roe Sash Or Sell\n\nMcMaster On Roe Ice Shop Or Stash\n\nMcMaster On Thunder Shell Or Sell\n\nMcMaster On Thunder Penalty Or Sell\n\nMcMaster . . .

Figure 4. GPT-2 Small Redundancy minimum outlier at -202.5

Human assessment of the text suggests that the Redundancy score requires improvement. Figure 4. presents an outlier for the generated text that has a higher score than the real text when the seeming redundancy is clearly the opposite.



```
                         Results: Logit
================================================================
Model:              Logit            Pseudo R-squared: 0.000
Dependent Variable: real             AIC:              2773.9649
Date:               2021-08-01 05:12 BIC:              2779.5658
No. Observations:   2000             Log-Likelihood:   -1386.0
Df Model:           0                LL-Null:          -1386.3
Df Residuals:       1999             LLR p-value:      nan
Converged:          1.0000           Scale:            1.0000
No. Iterations:     3.0000
------------------------------------------------------------------
           Coef.    Std.Err.     z      P>|z|     [0.025    0.975]
------------------------------------------------------------------
gruen      0.0664     0.0841   0.7897   0.4297   -0.0985    0.2314
================================================================
```



```
                          Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.014
Dependent Variable: real             AIC:              2739.7447
Date:               2021-08-01 05:13 BIC:              2756.5474
No. Observations:   2000             Log-Likelihood:   -1366.9
Df Model:           2                LL-Null:          -1386.3
Df Residuals:       1997             LLR p-value:      3.6740e-09
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
-----------------------------------------------------------------
                Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
-----------------------------------------------------------------
grammaticality  0.1577   0.0736  2.1417 0.0322  0.0134  0.3021
redundancy      0.0012   0.0015  0.8011 0.4231 -0.0018  0.0043
focus          -5.0069   1.1791 -4.2465 0.0000 -7.3178 -2.6960
=================================================================
```

Not only does logistic regression analysis of the component scores on a real-or-generated label indicate that Redundancy is likely insignificant as it is currently calculated, but so is the overall GRUEN score.

Conclusion and Future Work

The rationale behind the GRUEN metric is compelling. Combine and estimate important linguistic characteristics to better mimic how human reviewers may evaluate text. The authors suggest their metric has generalizable qualities, but it performs unreliably here. Developing how each component is scored may be fundamental to proving such an aggregation metric can work.

Improvements could be made in how the individual scores are combined into the GRUEN. Prior to simple summation, each component score could be assigned weights that are adaptive to the corpus's linguistic style and structure. Such a process might be possible while maintaining the unsupervised nature of the metric. Redundancy differs from corpus to corpus and can be intentional. The score should try to estimate this intention prior to final calculation. This may be possible if the individual scoring functions accounted for interactions between one another. A metric underpinned by a neural network architecture that can selectively pass information across the individual scorers might better capture quality. Further studies should incorporate comparisons with the other evaluation metrics as well as removal of the lower and upper GRUEN score bounds.

References

References

Asli Celikyilmaz, Elizabeth Clark and Jianfeng Gao. 2021. Evaluation of Text Generation: A Survey. *arXiv*, 006.14799(v2).

Eric Malmi, Pyry Takala, Hannu Toivonen, Tapani Raiko and Aristides Gionis. 2016. DopeLearning. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

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

Wanzheng Zhu and Suma Bhat. 2021. GRUEN for Evaluating Linguistic Quality of Generated Text. *arXiv*, 2010.02498v1.

Yusu Qian. 2020. Story-level Text Style Transfer: A Proposal. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*.