Is this the end of the gold standard? A straightforward reference-less grammatical error correction metric

Md Asadul Islam
mdasadul@ualberta.ca*

Enrico Magnani
enrico.magnani@scribendi.com

Abstract

It is difficult to rank and evaluate the performance of grammatical error correction (GEC) systems, as a sentence can be rewritten in numerous correct ways. A number of GEC metrics have been used to evaluate proposed GEC systems; however, each system relies on either a comparison with one or more reference texts—in what is known as the gold standard for reference-based metrics—or a separate annotated dataset to fine-tune the reference-less metric. Reference-based systems have a low correlation with human judgement, cannot capture all the ways in which a sentence can be corrected, and require substantial work to develop a test dataset. We propose a reference-less GEC evaluation system that is strongly correlated with human judgement, solves the issues related to the use of a reference, and does not need another annotated dataset for fine-tuning. The proposed system relies solely on commonly available tools. Additionally, currently available reference-less metrics do not work properly when part of a sentence is repeated as opposed to reference-based metrics. In our proposed system, we look to address issues inherent in reference-less metrics and reference-based metrics.

1 Introduction

Evaluating the performance of a machine translation, text summarization, text simplification or GEC system poses significant difficulties because there is more than one possible correct output (Choshen and Abend, 2018). Typically, a GEC system is evaluated by comparing changes made by the system with annotated gold standards called MaxMatch (M2) (Dahlmeier and Ng, 2012), I-measure (Felice and Briscoe, 2015), or Generalized Language Evaluation Understanding (GLEU) (Napoles et al., 2015). In each of these reference-based metrics, system outputs that differ from the annotated gold output are penalized. Additionally, some metrics use multiple references; the CoNLL-2014 (Ng et al., 2014) test set has 10 references that require significant time and effort to build, and they still might not cover all the possibilities. To demonstrate how a single sentence can be edited in many ways, let us consider the following sentence from the NUCLE-2014 test set (Macdonell, 2019):

> *Bigger farming are use more chemical product and substance to feed fish.*

One concise revision of this sentence would be as follows:

> *Big farms use more chemicals to feed fish.*

Likewise, if the author wished to express an ongoing action, the sentence could be the following:

> *Bigger farms are using more chemical products and substances to feed fish.*

We could go on. The point is that corrections to the raw data are not absolute because a change to one word in a sentence could alter another ostensibly erroneous word elsewhere in the sentence. Given the contextual nature of such corrections, human judgement is the best way to determine the quality of generated sentences (Grundkiewicz et al., 2015). In this paper, we propose the Scribendi Score, which is a reference-less metric to evaluate GEC system outputs that strongly correlates with human judgement. This paper makes three main contributions:

- We have determined that the "perplexity score" can be used to measure grammaticality and fluency. We also use two common metrics that

*This work was done when the author was a Machine Learning Engineer at Scribendi Inc.; currently he is Senior Machine Learning Engineer at People.ai

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3009–3015
November 7–11, 2021.

do not require further work, such as building a dataset to fine tune the metrics, and combine them into a single score that strongly correlates with human judgment.

- We have proposed a metric that performs excellently in comparison with reference-based metrics.
- We have developed and released code for the Scribendi Score that can be used to evaluate the output of a GEC system.

2 Related Work

Much work has addressed reference-based metrics, such as M2 (Dahlmeier and Ng, 2012), I-measure (Felice and Briscoe, 2015), GLEU (Napoles et al., 2015), and ERRANT (Korre and Pavlopoulos, 2020), but research on reference-less metrics is lacking. Napoles et al. (2016) used three grammaticality metrics to measure sentence quality. They used a proprietary e-rater grammatical error detection module and a language tool. They also used a linguistic feature-based model (Heilman et al., 2014) with ridge regression. Asano et al. (2017) used grammaticality, fluency, and meaning-preservation metrics to achieve better correlation with human judgement. Yoshimura et al. (2020) did the same for their sub-metrics that are optimized for manual evaluation (SOME). Grammaticality and fluency can be replaced by perplexity scores generated by a language model, such as the Generative Pretrained Transformer 2 (GPT-2) (Radford et al., 2019). To ensure the meaning is unaltered, we use the Levenshtein distance ratio and token sort ratio. Previous works have the following issues:

- Although (Napoles et al., 2016) used a reference-less metric, it relies on a proprietary system to detect grammatical errors.
- Manually annotated "Grammatical versus UnGrammatical" (GUG) (Heilman et al., 2014) and JHU FLuency-Extended GUG (JFLEG) (Napoles et al., 2017) corpora were used by Asano et al. (2017) and Yoshimura et al. (2020). Moreover, Yoshimura et al. (2020) used manually annotated datasets for further fine tuning.

It is difficult to collect and annotate GUG and JFLEG data or the data needed by SOME. Using annotated data and fine-tuning works well with the 12-reference system (Ng et al., 2014), but it may not work well with another dataset as the fine tuning might risk overfitting the test data.

3 Perplexity Score for Measuring Grammaticality

Language modeling is an approach to understanding linguistic structures by learning from a large corpus. Linguists have accumulated many corpora to find syntactic language rules and formalize them into standard grammar (Manning and Schutze, 1999). However, grammar cannot account for all situations in language use, as people sometimes speak ungrammatically in daily communication (Sapir, 1921). To address this problem, scientists use statistical modeling to identify common patterns within languages. Language models can learn the probability distribution of words in a sequence within a corpus. This approach is called language modeling. The language model is trained to minimize the cross-entropy loss, which is the same as minimizing perplexity. Recently, the perplexity score has been used to assess writing quality (Liu et al., 2020; Keukeleire, 2020).

Consider the following example: "Rarely read novels who reads comics." A GEC model might come up with the following correction: "Rarely read novels, who reads comics." But a human might suggest the following: "He/she who reads comics, rarely reads novels." Grundkiewicz et al. (2015) stated that the GEC system makes a small modification to the input sentences, which is why the outputs overlap significantly with the source sentences and the sentences produced by other systems. GEC systems struggle to suggest corrections that are longer than a few words, and they are also not good at reordering sentences when necessary. By using perplexity scores, we are trying to compare sentences with small modifications to the original sentences to determine whether the modification improves the grammaticality and fluency of the sentence.

4 Methodology

We use GPT-2 (Radford et al., 2019) without further fine tuning to measure the perplexity of a source and predictions from different models, and we use the output of the 12 systems evaluated in CoNLL-2014's shared tasks for GEC (Ng et al., 2014). The test set consists of 1312 sentences. For each predicted sentence that is the same as

Src 1: Once the test is done, whether the results should be open to his or her relatives has caused **social extensive** controversy.			
Pred 1: Once the test is done, whether the results should be open to his or her relatives has caused **extensive social** controversy.			
PPL Src 1	PPL Pred 1	TSR	LDR
104.48	62.72	100	94.308
Src 2: We **can not** let it go.			
Pred 2: We **cannot** let it go.			
PPL Src 2	PPL Pred 2	TSR	LDR
26.46	24.299	82.05	97.67

Table 1: A comparison between TSR: token sort ratio and LDR: Levenshtein distance ratio with their corresponding PPL: perplexity scores

its source, the score will be 0. We then calculate the perplexity score of each remaining output sentence of the 12 systems and compare the perplexity score with the perplexity score of its corresponding source sentence. If the perplexity score of the predicted sentence is greater than or equal to that of the source sentence, the Scribendi Score will be -1, as the perplexity score indicates that the suggested change(s) did not improve the grammaticality and fluency of the sentence. If the perplexity score improves (i.e., decreases), then we check whether the predicted sentences are syntactically related to the source sentence. We use the following two simple yet effective measurements to determine if they are syntactically similar as described at the end of Section 3: a token sort ratio and a Levenshtein distance ratio.

4.1 Token sort ratio

We use a token sort ratio to check the correspondence between the source and the prediction. A token sort ratio splits the sentences into tokens, sorts them, and finds the ratio of tokens that are the same between the two sorted sequences. This is helpful when sentences are reordered without being completely rewritten. The string "Fat Cat" and "Cat Fat" have a 100% match with the token sort ratio (Shah, 2019). We present a comparison between a token sort ratio and a Levenshtein distance ratio. The first example in Table 1 shows the effectiveness of the token sort ratio. If this ratio is high ($\geq 80\%$), we consider the change to be good.

4.2 Levenshtein distance ratio

The Levenshtein distance (Levenshtein, 1965) is calculated between the source and the prediction. We consider the cost of insertion or deletion to be 1 and the cost of replacement to be 2. From the Levenshtein distance (LD), we calculate the Levenshtein distance ratio (LDR) as follows:

$$LDR = 1 - \frac{LD}{len(source) + len(prediction)}.$$

where $len(source)$ and $len(prediction)$ indicate the number of characters in the source and prediction sentences. A ratio higher than 80%, which is chosen empirically, is a good indicator that the source and prediction sentences are similar. The second example in Table 1 shows that the Levenshtein distance score can effectively measure similarities between two sentences even when the token sort ratio is low.

Listing 1: Scribendi Score

```
def Scribendi_score(src, pred):
    if pred == src:
        return 0
    ppl_source = Perplexity(src)
    ppl_predition = Perplexity(pred)
    if ppl_source<=ppl_prediction:
        return -1
    else:
        tsr = token_sort_ratio(src,pred)
        ldr = lev_dist_ratio(src,pred)
        if max(tsr, ldr)>=0.8: return 1
        else: return -1
```

We find the maximum value between the token sort ratio and the LDR. If the score is $\geq 80\%$, then we assume that the overall meaning of the sentence has not changed. Otherwise, we consider the predicted sentence to be unrelated to the source sentence and will mark it as a poor change. According to this approach, a good change is scored as $+1$. If the source and prediction are the same, we assign a score of 0; otherwise, we assign a score of -1 to reflect that the correction does not improve the sentence. Finally, we calculate the system score for a particular system by summing up the individual sentence scores. Listing 1 shows the pseudocode of the Scribendi metric.

5 Results and Discussion

Following the CoNLL-2014 shared task on grammatical error correction (Ng et al., 2014), for which all the results are publicly available, including the references and 12 system outputs, Grundkiewicz

et al. (2015) and Napoles et al. (2015) simultaneously performed a human evaluation of the system outputs. In this section, we present Tables 2, 3, and 4, which compare the Scribendi Scores to the corresponding human evaluations. In Table 5 in Appendix A, we took the ranking of different systems from best to worst according to the metrics presented in Napoles et al. (2015) and added the Scribendi Score to it. The human-generated ranking differs significantly from all the reference-based metrics. We can also see that the source is ranked somewhere in the middle according to the human evaluation, the Scribendi Score, and the GLEU scores. If no correction is necessary then the source sentence should be the best possible choice. Some models (e.g., IPN) also introduce errors in their attempted corrections, which results in a lower ranking. As Napoles et al. (2015) noted, the human-to-human Pearson correlation for ranking the 12 different system outputs is $0.73 \leq r \leq 0.81$. In Table 2, the Scribendi Score is 0.780 for Pearson and 0.812 for Spearman, which is significantly higher than the GLEU score for this task and is similar to the heavily fine-tuned SOME metric. From the results presented below, we can conclude that the reference-less metric is significantly correlated with human evaluation in comparison with reference-based metrics.

	Pearson	**Spearman**
M2	0.429	0.358
GLEU	0.555	0.542
Scribendi Score	0.780	0.812
SOME	**0.824**	**0.824**

Table 2: Pearson and Spearman's correlation of metrics with human ranking from Napoles et al. (2016). We calculated the Person and Spearman scores for SOME.

Grundkiewicz et al. (2015) used extensive measurements and computed the expected human-generated ranking and the Human TrueSkill ranking. Table 3 is based on the Human TrueSkill ranking. It is clear that the Scribendi Score performs competitively in comparison with finely tuned measurements, such as SOME and those used by Asano et al. (2017).

Current state-of-the-art GEC models are based on neural language models (Omelianchuk et al., 2020; Kaneko et al., 2020). Neural language models are well known for generating repeated words (See et al., 2019; Dathathri et al., 2019). Table 4 shows an example of repeated words and the scores according to various metrics. Reference-based metrics are able to capture this kind of repetition, but reference-less metrics are unable to address it since they use language models that assign better scores to such sentences. This phenomenon has also been reported by Yoshimura et al. (2020), who show that reference-based metrics are still quite useful for addressing this problem. We use the Levenshtein distance to address this problem, which is capable of capturing the repetition in this kind of situation and solving this problem of SOME (Yoshimura et al., 2020).

	Pearson	**Spearman**
M2	0.674	0.720
GLEU	0.846	0.816
Asano et al. (2017)	0.878	0.874
Scribendi Score	0.951	0.940
SOME	**0.975**	**0.978**

Table 3: Pearson and Spearman's correlation between metrics and human-generated rankings from Grundkiewicz et al. (2015)

Source:	He is going school.
Reference:	He is going to school.
Prediction:	He He He He He He.

Manual Eval	M2	GLEU	SOME	**Scribendi Score**
X	0.37	0.22	0.87	**-1**

Table 4: The weakness of SOME (Yoshimura et al., 2020)

There were two main reasons that pushed us to use a discrete score. First, perplexity scores can vary greatly between sentences. Let us consider two pairs of sentences. Please note that a lower perplexity score is better.

Source: People get certain disease because of genetic changes. Perplexity Score: 148.57 Target: People get certain diseases because of genetic changes. Perplexity Score: 80.62

Source: The basis of a family is that everyone trusts and love each other with no doubts. Perplexity Score: 52.92 Target: The basis of a family is that everyone trusts and loves each other with no doubts. Perplexity Score: 44.09

We can see that the perplexity scores of the two examples above are different despite the applica-

tion of a similar correction. Also note that the perplexity score variation from source to target differs greatly between the two examples. Second, if we think about the task from a human perspective, there is no objective metric for gauging the importance of certain corrections within a sentence. For example, is a tense correction more important than subject-verb agreement? Is a word order issue less important than a spelling mistake? What happens if there are multiple mistakes and different ways to correct the sentence? On the other hand, we can easily define whether or not a sentence has improved. We just need to show that the target sentence has no or fewer mistakes than the source sentence while maintaining the meaning. By discretizing the Scribendi Score, we ensure that the target sentences in the above examples both have the same score of 1, which means that the target sentences are more grammatically correct than the source sentences.

We initially considered combining the perplexity score with the token sort ratio and Levenshtein score, but found that it does not work in situations where one of those scores is significantly lower than the other. Such cases would make the combined score low even when it is a good change. Let us consider the following sentences:

Source: More and more illness are discovered to be related to some genes with the development of the medical technology. Perplexity Score: 81.93 Target: With the development of medical technology, more and more illnesses have been discovered to be related to some genes. Perplexity Score: 20.232

The Levenshtein Score and token sort ratio output a number between 0 and 1; 0 if the sentences are totally different and 1 if they are exactly the same. In the example above, we can see that the target is better than the source sentence, although the Levenshtein score (0.554) is significantly lower than the token sort ratio (0.929). Combining the two scores in this case could generate the wrong result.

6 Summary

In this paper, we identified the shortcomings of reference-less metrics. Namely, they need another annotated dataset for fine tuning and do not work properly when part of a sentence is repeated. We also highlighted that reference-based metrics are unable to score a sentence properly when the predicted corrections are not contained in the reference sentences. In this study, we evaluated source and system outputs using the Scribendi Score, which is based on the perplexity score, the token sort ratio, and the Levenshtein distance ratio. We demonstrated that the Scribendi Score does not require an extra annotated dataset for fine tuning, which is expensive in terms of resources and could cause over-fitting of certain datasets. It is strongly correlated with human evaluation, and is able to address the issue of repetition.

7 Acknowledgements

We would like to thank Pascal Poupart (Professor in the David R. Cheriton School of Computer Science at the University of Waterloo) and Gene Saunders (Machine Learning Engineer at Scribendi Inc.) for their support throughout the research work. The completion of this work could not have been possible without the extensive support of the whole team at Scribendi Inc., especially the many in-house editors who have contributed to the research and edited and proofread this paper.

References

Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. 2017. Reference-based metrics can be replaced with reference-less metrics in evaluating grammatical error correction systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–348.

Leshem Choshen and Omri Abend. 2018. Reference-less measure of faithfulness for grammatical error correction. *arXiv preprint arXiv:1804.03824*.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Mariano Felice and Ted Briscoe. 2015. Towards a standard evaluation method for grammatical error detection and correction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 578–587.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. 2015. Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180.

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. *arXiv preprint arXiv:2005.00987*.

Pia Keukeleire. 2020. Correspondence between perplexity scores and human evaluation of generated tv-show scripts.

Katerina Korre and John Pavlopoulos. 2020. Errant: Assessing and improving grammatical error type classification. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 85–89.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10.

Yang Liu et al. 2020. Assessing text readability and quality with language models.

Cameron Macdonell. 2019. Grammatical error correction tools a novel method for evaluation. (accessed 10-April-2021).

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. 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)*, pages 588–593.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. *arXiv preprint arXiv:1610.02124*.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. Gector–grammatical error correction: Tag, not rewrite. *arXiv preprint arXiv:2005.12592*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Edward Sapir. 1921. *An introduction to the study of speech*. Citeseer.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. 2019. Do massively pretrained language models make better storytellers? *arXiv preprint arXiv:1909.10705*.

Darshak Shah. 2019. Fuzzy string matching in python. (accessed 10-April-2021).

Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. Some: Reference-less sub-metrics optimized for manual evaluations of grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6516–6522.

A Appendix

Ranking	Human	BLEU	I-measure	M2	GLEU	Scribendi Score	SOME
1	CAMB	UFC	UFC	CUUI	CUUI	CAMB	AMU
2	AMU	Source	Source	CAMB	AMU	AMU	CAMB
3	RAC	IITB	IITB	AMU	UFC	CUUI	RAC
4	CUUI	SJTU	SJTU	POST	CAMB	RAC	POST
5	Source	UMC	CUUI	UMC	Source	PKU	CUUI
6	POST	CUUI	PKU	NTHU	IITB	UMC	UMC
7	UFC	PKU	AMU	PKU	SJTU	POST	PKU
8	SJTU	AMU	UMC	RAC	PKU	Source	Source
9	IITB	IPN	IPN	SJTU	UMC	IITB	UFC
10	PKU	NTHU	POST	UFC	NTHU	UFC	IITB
11	UMC	CAMB	RAC	IPN	POST	SJTU	SJTU
12	NTHU	RAC	CAMB	IITB	RAC	NTHU	NTHU
13	IPN	POST	NTHU	Source	IPN	IPN	IPN

Table 5: Ranking of source sentences and 12 systems by different metrics from Napoles et al. (2015)