A Reassessment of Reference-Based Grammatical Error Correction Metrics

Shamil Chollampatt[1] and **Hwee Tou Ng**[1,2]
[1]NUS Graduate School for Integrative Sciences and Engineering
[2]Department of Computer Science, School of Computing
National University of Singapore
shamil@u.nus.edu,nght@comp.nus.edu.sg

Abstract

Several metrics have been proposed for evaluating grammatical error correction (GEC) systems based on grammaticality, fluency, and adequacy of the output sentences. Previous studies of the correlation of these metrics with human quality judgments were inconclusive, due to the lack of appropriate significance tests, discrepancies in the methods, and choice of datasets used. In this paper, we re-evaluate reference-based GEC metrics by measuring the system-level correlations with humans on a large dataset of human judgments of GEC outputs, and by properly conducting statistical significance tests. Our results show no significant advantage of GLEU over MaxMatch (M^2), contradicting previous studies that claim GLEU to be superior. For a finer-grained analysis, we additionally evaluate these metrics for their agreement with human judgments at the sentence level. Our sentence-level analysis indicates that comparing GLEU and M^2, one metric may be more useful than the other depending on the scenario. We further qualitatively analyze these metrics and our findings show that apart from being less interpretable and non-deterministic, GLEU also produces counter-intuitive scores in commonly occurring test examples.

1 Introduction

Grammatical error correction (GEC) refers to the task of automatically detecting and correcting grammatical, spelling, and word choice errors in written text. As newer approaches are being developed for this task, it becomes essential to have reliable means to compare them. In the related field of machine translation (MT), human judgments are used as ground truth to rank systems in the evaluation campaigns as part of the Workshop on Machine Translation (WMT) (Bojar et al., 2016a) held annually. In the absence of annual evaluations and periodic human-judged shared tasks for GEC, robust automatic evaluation metrics are necessary to reliably assess improvements. Automatic evaluation methods can also help in system development and validation.

Previous shared tasks on GEC relied on automatic evaluation metrics to evaluate the performance of participating systems (Dale and Kilgarriff, 2011; Dale et al., 2012; Ng et al., 2013; Ng et al., 2014; Rozovskaya et al., 2015). The CoNLL-2014 shared task test set and evaluation metric, M^2 (Dahlmeier and Ng, 2012), have been used as the primary benchmark for evaluating English GEC. After the CoNLL-2014 shared task, two reference-based evaluation metrics, namely I-measure (Felice and Briscoe, 2015) and GLEU (Napoles et al., 2015; Napoles et al., 2016a), were proposed for GEC. More recently, some reference-less metrics were also proposed (Napoles et al., 2016b; Asano et al., 2017). A standard way to identify the best metric among them is to compare their correlation to human judgments. In the case of MT, WMT organizes a metrics shared task each year, where the correlation of system-level rankings generated by metrics and humans is measured. There is also a segment-level evaluation subtask where the agreement of metrics is measured at the sentence level. Prior work in GEC has followed the system-level evaluation approach to compare metrics (Grundkiewicz et al., 2015; Napoles et al., 2015; Sakaguchi et al., 2016). However, there has been no prior work on sentence-level evaluation. From prior studies,

Proceedings of the 27th International Conference on Computational Linguistics, pages 2730–2741
Santa Fe, New Mexico, USA, August 20-26, 2018.

it may appear that GLEU performs better than the de-facto M^2 metric. However, previous system-level correlation studies turn out to be inadequate, primarily due to the absence of significance tests and the choice of datasets and methods used. In this paper, we re-evaluate reference-based GEC metrics using appropriate significance tests to measure if previous conclusions can be trusted. We find that there is no evidence to suggest that GLEU is significantly better than M^2. Contrary to system-level evaluation, I-measure is found to be a reasonably useful metric at the sentence level and has a positive correlation with human judgments at the sentence level. In our sentence-level evaluation, we find scenarios where M^2 outperforms GLEU and vice versa. Our qualitative assessment of these metrics on example sentences reveals the shortcomings of GLEU in comparison to the other metrics.

2 Evaluation Metrics

We study three popular reference-based evaluation measures that have been proposed for GEC, namely, MaxMatch (M^2), I-measure, and GLEU.

2.1 MaxMatch (M^2)

The M^2 metric (Dahlmeier and Ng, 2012) computes precision, recall, and F-measure by maximally matching phrase-level edits made by a system to gold-standard edits annotated by humans. A gold-standard edit used by M^2 includes the location of the error within the tokenized input sentence and the proposed correction. Although annotations about the type of errors can be included in the gold standard, they are not used while scoring. For computing scores for specific error types, a variant of M^2 has been proposed (Bryant et al., 2017).

The standard M^2 computation is defined as follows. Consider a set of input sentences $S = \{s_1, ..., s_n\}$ and their corresponding system corrected hypotheses $H = \{h_1, ..., h_n\}$. The set of gold-standard edits for each input sentence s_i is denoted by $\mathbf{g}_i$ and the set of edits made by the system to transform s_i to h_i is denoted by $\mathbf{e}_i$. Precision, recall, and F-measure are given by:

$$\text{precision} = \frac{\text{No. of correct edits made by the system}}{\text{Total no. of edits made by the system}} = \frac{\sum_{i=1}^{n} |\mathbf{e}_i \cap \mathbf{g}_i|}{\sum_{i=1}^{n} |\mathbf{e}_i|}$$

$$\text{recall} = \frac{\text{No. of correct edits made by the system}}{\text{Total no. of gold-standard edits}} = \frac{\sum_{i=1}^{n} |\mathbf{e}_i \cap \mathbf{g}_i|}{\sum_{i=1}^{n} |\mathbf{g}_i|}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}.$$

A β value of 0.5 is used in standard M^2 (since CoNLL-2014 shared task), in order to weight precision twice as much as recall. This penalizes incorrect feedback more severely, which is especially important in the context of language learning, where we would like to minimize giving incorrect feedback to language learners. When several sets of annotations (by different annotators) are available for an input sentence s_i, the set that maximizes the F-measure is chosen as $\mathbf{g}_i$.

The set of system edits $\mathbf{e}_i$ for the ith sentence is obtained by first constructing a token-level edit lattice from the Levenshtein distance matrix. Additional edges are added to the lattice to represent phrase-level edits by combining adjacent edges, subject to a constraint on the maximum number of unchanged words (set to 2 in standard M^2). Edges of the lattice are appropriately weighted such that a minimum distance path computation yields the set of edits $\mathbf{e}_i$ that maximally overlaps with the gold-standard edits.

2.2 I-measure

I-measure (Felice and Briscoe, 2015) is a token-level accuracy-based metric proposed to address the inability of M^2 to distinguish between a system that does not correct any errors and a system that only makes wrong changes. Annotations for I-measure can involve non-contiguous tokens as well. The computation of I-measure begins by first aligning the input, hypothesis, and reference corrections in a three-way token-level alignment. Gaps in the alignment can be assumed to be marked by NULL (ϵ) tokens at appropriate positions (Table 1). Let the aligned input, hypothesis, and reference tokens at

POSITIONS	1	2	3	4	5	6	7	8	9	10	11	12	13
INPUT:	Thus	,	advice	from	ϵ	hospital	plays	the	important	role	for	this	.
HYPOTHESIS:	Thus	,	advice	from	ϵ	hospital	plays	an	important	role	for	this	.
REFERENCE:	Thus	,	advice	from	the	hospital	plays	an	important	role	in	this	.

Table 1: An example of three-way token-level alignments produced by I-measure.

position j be denoted by w_j^{inp}, w_j^{hyp}, and w_j^{ref}, respectively. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are defined as follows:

$$\text{TP}: w_j^{\text{inp}} \neq w_j^{\text{ref}} \text{ and } w_j^{\text{hyp}} = w_j^{\text{ref}} \qquad \text{FP}: w_j^{\text{inp}} \neq w_j^{\text{hyp}} \text{ and } w_j^{\text{hyp}} \neq w_j^{\text{ref}}$$

$$\text{TN}: w_j^{\text{inp}} = w_j^{\text{ref}} = w_j^{\text{hyp}} \qquad \text{FN}: w_j^{\text{inp}} \neq w_j^{\text{ref}} \text{ and } w_j^{\text{hyp}} \neq w_j^{\text{ref}}$$

An additional class FPN is defined to balance cases that can be classified as both FP and FN ($\text{FPN}: w_j^{\text{inp}} \neq w_j^{\text{ref}} \neq w_j^{\text{hyp}}$). A weighted accuracy is computed using TP, FP, TN, and FN counts:

$$\text{WAcc} = \frac{\lambda \cdot \text{TP} + \text{TN}}{\lambda \cdot \text{TP} + \text{TN} + \lambda \cdot \left(\text{FP} - \frac{\text{FPN}}{2}\right) + \left(\text{FN} - \frac{\text{FPN}}{2}\right)}$$

The weight λ, which is set to 2, rewards correct changes and penalizes incorrect changes more than preserving erroneous input tokens. When multiple references are used, similar to M^2, the gold-standard reference that maximizes the WAcc is chosen. Then, the weighted accuracy of the input (WAcc_{inp}) is computed by considering the input sentences as the hypothesis, with the same set of references chosen to compute WAcc. I-measure is given by,

$$\text{I} = \begin{cases} \lfloor \text{WAcc} \rfloor, & \text{if WAcc} = \text{WAcc}_{\text{inp}} \\ \dfrac{\text{WAcc} - \text{WAcc}_{\text{inp}}}{1 - \text{WAcc}_{\text{inp}}}, & \text{if WAcc} > \text{WAcc}_{\text{inp}} \\ \dfrac{\text{WAcc}}{\text{WAcc}_{\text{inp}}} - 1, & \text{otherwise.} \end{cases}$$

I-measure denotes the relative improvement or degradation with respect to the input text. I-measure falls in the range[1] $[-1, 1]$, a negative value indicates degradation and a positive value indicates improvement. Unlike M^2, I measure can mix and match annotations from different annotators to produce more alternative references[2].

2.3 GLEU

Unlike M^2 and I-measure, GLEU (Napoles et al., 2015; Napoles et al., 2016a) only requires human annotators to correct by re-writing the source sentence without requiring annotations for individual errors. GLEU computes the precision of n-grams in the hypothesis that match part of the reference sentence, similar to the MT metric BLEU (Papineni et al., 2002). Additionally, GLEU penalizes n-grams in the hypotheses that match part of the input but not the reference. The original formulation (Napoles et al., 2015) included a weight parameter that had to be re-tuned according to the number of reference corrections used. Following the recommendation of Napoles et al. (2016a), we use the new formulation of GLEU[3] which does not include this weight parameter and can work for any number of references.

The computation of GLEU is done as follows. Consider a set of input sentences $S = \{s_1, ..., s_n\}$, their corresponding corrected hypotheses $H = \{h_1, ..., h_n\}$, and reference sentences, $R = \{r_1, ..., r_n\}$.

[1] I-measure may also be represented as a percentage (%) with a range $[-100, 100]$ as used in the rest of the paper.

[2] To use this ability of mixing annotations, alternative corrections for the same underlying error must be grouped together during annotation. However, since the data we use was not annotated in this manner, we disable the mixing ability of I-measure by using `-nomix` to prevent generating invalid references.

[3] This is referred to as GLEU+ in (Napoles et al., 2016a).

Here, we assume that there is a single reference sentence for each input sentence. First, a precision term p_k is computed for n-grams of size k ($k = 1, 2, ..., N$ and $N = 4$ for standard GLEU):

$$p_k = \frac{\sum\limits_{h_i \in H} \left(\sum\limits_{\substack{\text{k-gram in} \\ h_i \text{ and } r_i}} \text{count}_{h_i,r_i}(\text{k-gram}) - \sum\limits_{\substack{\text{k-gram in} \\ h_i \text{ and } s_i}} \max\left[0, \text{count}_{h_i,s_i}(\text{k-gram}) - \text{count}_{h_i,r_i}(\text{k-gram})\right] \right)}{\sum\limits_{h_i \in H} \sum\limits_{\substack{\text{k-gram} \\ \text{in } h_i}} \text{count}_{h_i}(\text{k-gram})}$$

$$\text{count}_a(\text{n-gram}) = \#\text{ occurrences of n-gram in } a$$
$$\text{count}_{a,b}(\text{n-gram}) = \min(\#\text{occurrences of n-gram in } a, \#\text{occurrences of n-gram in } b)$$

Similar to BLEU, a brevity penalty (BP) is computed to penalize short hypotheses:

$$\text{BP} = \begin{cases} 1, & \text{if } l_h > l_r \\ \exp(1 - l_r/l_h), & \text{if } l_h \leq l_r \end{cases}$$

where l_r is the reference corpus length (sum of the number of tokens of the reference sentences) and l_h is the total hypothesis corpus length (sum of the number of tokens of all hypotheses). GLEU is given by:

$$\text{GLEU(S, H, R)} = \text{BP} \cdot \exp\left(\frac{1}{N} \sum_{k=1}^{N} \log p_k \right)$$

When multiple references are available, GLEU does not include reference n-grams from all references to compute n-gram precision like BLEU, nor picks the best like M^2 and I-measure. Instead, for each input sentence, a random reference correction is chosen from the set of reference corrections to compute the GLEU score. This makes GLEU non-deterministic, unlike M^2 and I-measure. The average GLEU score over m GLEU score computations is reported as the final score ($m = 500$ in standard GLEU).

3 Quantitative Evaluation

The three GEC metrics are evaluated by measuring their correlation with human quality judgments treated as the ground truth. Following the experimental methodology in WMT metrics shared tasks (Macháček and Bojar, 2014; Stanojević et al., 2015; Bojar et al., 2016b; Bojar et al., 2017), we evaluate system-level correlation as well as sentence-level agreement of metrics with human judgments.

We utilize the collection of human judgments of GEC outputs released in (Grundkiewicz et al., 2015). The dataset is annotated similar to the relative-ranking method adopted in the WMT metrics shared tasks in (Macháček and Bojar, 2014; Stanojević et al., 2015). Human judgments are obtained for the system outputs of the 12 participating systems from the CoNLL-2014 shared task (Ng et al., 2014) and the input text as the thirteenth system (referred to as INPUT). Each human judgment consists of a 5-way ranking of hypotheses corrections from randomly chosen systems for an input sentence. Systems that produce the same hypothesis for an input sentence are grouped together. The 5-way rankings are then converted to pairwise rankings that include ties. A total of 109,098 pairwise rankings can be obtained if systems with the same hypothesis that were grouped together in the 5-way rankings are included (henceforth, referred to as *expanded set*). Grundkiewicz et al. (2015) observed that there were a large number of ties due to the high overlap in system outputs. Trivial ties of systems that produce the same hypothesis can be removed by including only one randomly chosen system among grouped systems. This results in 20,516 pairwise rankings (henceforth, referred to as *unexpanded set*).

3.1 Measuring System-Level Correlation

System-level correlation is computed by comparing the ranking of the participating systems by humans and the ranking generated by the metric scores. Generating a ranking using metric scores is straightforward. However, human pairwise comparisons need to be converted into a ranking of systems. To do this, two methods are employed in (Grundkiewicz et al., 2015)[4], namely *Expected Wins* (Bojar et al., 2013) and *Trueskill* (Sakaguchi et al., 2014).

The Expected Wins model computes an intuitive score for a system based on the probability that a system wins (ranks higher according to human judgments) over a randomly chosen system on a randomly chosen input sentence. This model ignores ties. The Trueskill model computes a score representing the average relative ability of the system compared to other systems, accounting for the uncertainty surrounding the system's ability as well. In the context of GEC, on this dataset, Grundkiewicz et al. (2015) empirically found that Expected Wins performs better than Trueskill. However, almost all subsequent work ignored this finding and used the Trueskill model instead. We compute correlations using both methods in order to ensure comparability to prior work.

We measure both Pearson and Spearman correlation coefficients between metric rankings and human rankings. Pearson correlation (r) compares the system-level scores produced by a metric against human, under the assumption that the two measured variables have a linear relationship:

$$r = \frac{\sum\limits_{i=1}^{q}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{q}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{q}(y_i - \bar{y})^2}}$$

where $x_1, ..., x_q$ are metric scores for q systems for a particular metric with $\bar{x}$ as the mean metric score, and $y_1, ..., y_q$ are human scores computed by Trueskill or Expected Wins with $\bar{y}$ as the mean human score. On the other hand, Spearman correlation (ρ) computes a correlation coefficient based on ranks instead of scores:

$$\rho = 1 - \frac{6\sum\limits_{i=1}^{q}(d_i)^2}{q(q^2-1)}$$

where d_i represents the difference in the human rank and metric rank of the ith system. Spearman correlation is more relaxed compared to Pearson correlation in terms of the assumptions made about variables and also less sensitive to outliers in the sample.

In order to determine if a metric outperforms another, it is inadequate to measure differences in correlation alone. Following the recommendations in (Graham and Baldwin, 2014), we evaluate for significance of differences of correlation between metrics using William's test (Williams, 1959). Note that prior work in human evaluation of GEC systems has not reported significance tests that account for the dependence between two metrics and hence the derived conclusions are not justified.

3.2 Measuring Sentence-Level Agreement

Since system-level evaluation is done on a few systems (13 in our case), we also compute a fine-grained sentence-level agreement of metrics to human pairwise rankings. For this, sentence-level metric scores are computed[5]. Agreement to humans is computed in a similar manner as segment-level evaluation in the WMT metrics shared task and variants of the Kendall's Tau (τ) are used:

$$\tau = \frac{|\text{Concordant}| - |\text{Discordant}|}{\text{Total \# Pairwise Comparisons}} \tag{1}$$

where |Concordant| refers to the number of times a metric agrees with the sentence-level human pairwise comparisons, and |Discordant| refers to the number of times they disagree. The variants of τ are related to the way in which human ties are handled. In the variant used in the WMT metrics shared task from WMT14 (Macháček and Bojar, 2014) onwards (henceforth referred to as *NoTies*), human ties are ignored completely and metric ties are added to the denominator alone, without contributing to Concordant or Discordant sets. For GEC, since there is a large number of ties, particularly in the expanded set, we include another variant that incorporates human ties as well (henceforth, referred to as *HTies*). In this

[4] We use the scripts released by Grundkiewicz et al. (2015) to compute human ranking of systems.

[5] Similar to BLEU, sentence-level GLEU needs smoothing to avoid zero n-gram counts. We use the sentence-level scores produced by GLEU using the `-d` flag, where smoothing is performed by replacing a zero count with one.

Metric	Expected Wins		Trueskill	
	r	ρ	r	ρ
GLEU	0.691	0.407	0.733	0.478
I-MEASURE	–0.250	–0.385	–0.316	–0.423
M^2	0.623	0.687	0.672	0.720

Table 2: Results of system-level Pearson (r) and Spearman (ρ) correlations of GEC metrics using the Expected Wins model and Trueskill model of human rankings.

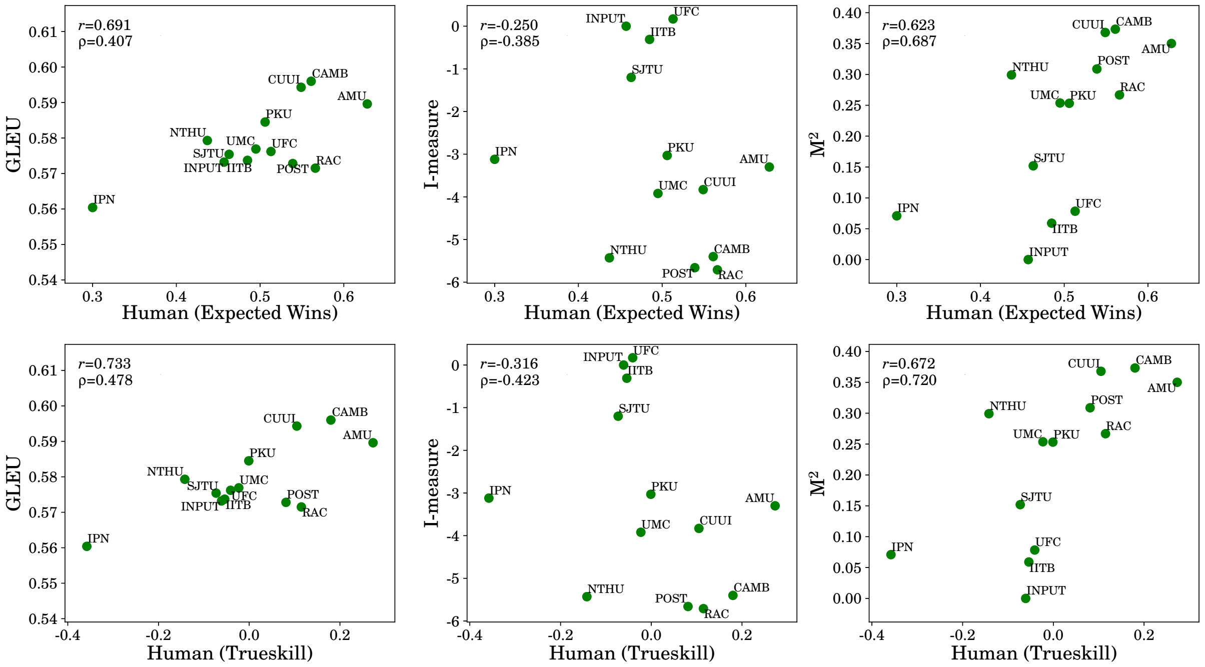


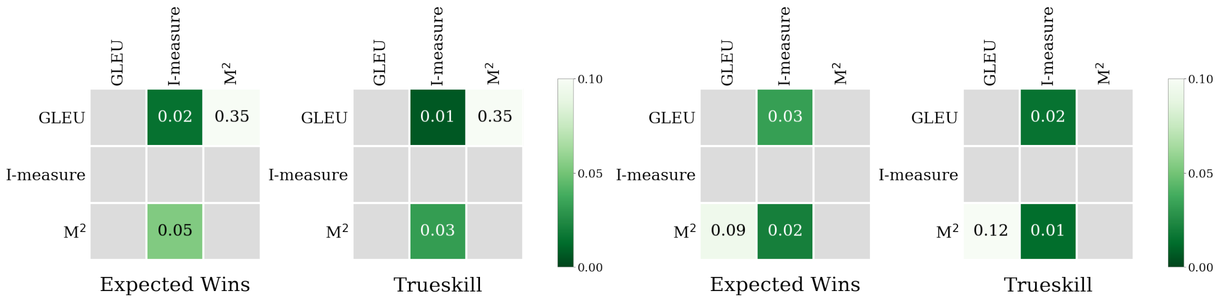
Figure 1: Scatter plots showing the metric scores and human scores of CoNLL-2014 shared task systems.

variant, when a human judges a tie and metric predicts a tie as well for a given pair of hypotheses, it is treated as a concordant pair. However, when a human predicts a tie and a metric does not, or vice versa, it contributes only to the denominator and not to the Discordant set. We test for significance similar to (Bojar et al., 2017), by employing bootstrap resampling over 1,000 samples and those metrics with non-overlapping 95% confidence intervals are treated as having statistically significant improvements compared to the lower performing metrics.

3.3 Results

3.3.1 System-Level Evaluation

The results of the system-level correlation tests are given in Table 2. In both Expected Wins and Trueskill methods of human ranking, GLEU achieves a higher Pearson correlation compared to the rest. M^2 achieves moderately high Pearson correlations and the highest Spearman correlations using both methods of human ranking. As noted in prior studies (Grundkiewicz et al., 2015; Napoles et al., 2016a), I-measure achieves a negative correlation. In order to decide which correlation measure is more suitable and to understand the cause of disagreement in both correlation measures, it may be worth looking at the bivariate scatter plot of metric and human scores (Figure 1). The presence of an outlier (the leftmost point – IPN) indicates that Pearson correlation may not be the ideal choice as it is sensitive to outliers in the data. To demonstrate this, we remove the system IPN from the ranking. Pearson correlation of GLEU then becomes lower than that of M^2 (0.500 for GLEU vs 0.593 for M^2 using Expected Wins and



(a) p-values for Pearson correlation.

(b) p-values for Spearman correlation

Figure 2: Results of William's significance tests for correlations between metrics. A green non-empty cell denotes that the row metric has a higher correlation coefficient compared to the column metric with the specified p-value. A gray empty cell indicates that the row metric has equal or lower correlation coefficient compared to the column metric.

0.584 for GLEU vs 0.638 for M^2 using Trueskill). The relative performance under Spearman correlation, however, remains the same (0.245 for GLEU vs 0.643 for M^2 using Expected Wins and 0.336 for GLEU vs 0.692 for M^2 using Trueskill). The high Spearman correlation for M^2 indicates that it is a good metric for ranking systems, which is the primary goal of automatic evaluation metrics.

To understand if one metric significantly outperforms another, we perform William's significance tests (Figure 2). As expected, we find that GLEU is not significantly better than M^2 in terms of Pearson correlation (p-value of 0.35 for both Expected Wins and Trueskill). In the case of Spearman correlation for Expected Wins, M^2 is significantly better than GLEU at the level of significance 0.10, but does not significantly outperform GLEU for the Trueskill model. The significance test results show that neither GLEU nor M^2 can be conclusively shown to be better than the other.

3.3.2 Sentence-Level Evaluation

Table 3 shows the results of sentence-level agreement of metrics to human pairwise rankings when human ties are considered (HTies) or ignored (NoTies). We also consider the unexpanded set where systems that produce the same hypothesis for an input sentence are grouped together. This removes the trivial ties (two identical hypotheses judged to be a tie) from the pairwise rankings. In the HTies variant, M^2 achieves statistically significant improvements compared to the other two metrics in both expanded and unexpanded sets. However, for NoTies variant, GLEU achieves the best result. The results show that M^2 may be better at judging ties and performs well overall when complete pairwise rankings are considered. On the other hand, GLEU does a better job at discriminating hypotheses, likely due to its ability to judge fluency like humans do (Sakaguchi et al., 2016). This difference in behavior of GLEU and M^2 between the two variants, HTies and NoTies, is expected given a large number of ties for GEC (Grundkiewicz et al., 2015) as indicated by the difference in the number of pairwise comparisons (Table 3) and that M^2 generally assigns more ties compared to GLEU. Surprisingly, I-measure achieves a positive correlation at the sentence level as opposed to a negative correlation at the system level, suggesting that it can be a useful metric at the sentence level.

4 Qualitative Assessment

We illustrate the strengths and weaknesses of the three metrics with the help of two examples (Table 4). We compare the metrics based on two criteria, interpretability and intuitiveness, which we believe are necessary for good GEC metrics.

Interpretability: The scores produced by a good GEC metric should be interpretable as a measure of a system's ability to correct errors and improve the text. M^2 and I-measure are based on minimal annotations and their results are interpretable. For example, M^2 measures the precision and recall of a system, in terms of the number of phrase-level errors that it corrects in a given input text. Similarly, I-measure is computed based on the relative weighted accuracy between a system and a do-nothing

Metric	Expanded		Unexpanded	
	HTies (109098)	NoTies (49981)	HTies (20516)	NoTies (14584)
GLEU	0.567	0.388*	0.237	0.321
I-MEASURE	0.564	0.368	0.242	0.293
M^2	0.617*	0.300	0.348*	0.266

Table 3: Results of sentence-level agreement in terms of two variations of Kendall's τ (HTies and NoTies) on expanded and unexpanded sets. The number of human pairwise comparisons used is given in parentheses. * indicates that the 95% confidence interval of the metric does not overlap with those of the other metrics.

		(GLEU, I%, M^2)
	Example 1	
INP:	The weekly quizzes in this course makes it challenging and fun .	
HYP1:	The weekly quizzes in this course makes it challenging and fun .	(0.392, 0.00, 0.000)
HYP2:	The weekly quizzes in this course ***making*** it challenging and fun .	(0.735, –4.00, 0.000)
REF:	The weekly quizzes in this course ***make*** it challenging and fun .	
	Example 2	
INP:	The senior student who failed have to retake the course next year .	
HYP1:	The senior student who failed ***has*** to retake the course next year .	(0.661, 100.00, 1.000)
HYP2:	The senior ***students*** who failed have to retake the course next year .	(0.656, 100.00, 1.000)
HYP3:	The senior ***students*** who failed ***has*** to retake the course next year .	(0.776, –6.11, 0.556)
REF1:	The senior student who failed ***has*** to retake the course next year .	
REF2:	The senior ***students*** who failed have to retake the course next year .	

Table 4: Illustrating examples and scores produced by the metrics.

baseline. In contrast, GLEU measures the precision of n-grams of the output text similar to BLEU. The scores that it produces have no clear relation to the system's ability to correct or improve a text. For example, in Example 1, Hypothesis 1 (Table 4), when the system hypothesis is exactly the input sentence itself, there is no evidence about the system's ability to correct errors. While I and M^2 give a zero score, GLEU gives a non-zero score (GLEU = 0.392). In the case of its counterpart BLEU, if some n-grams of the hypothesis and reference match, it rightly assigns a non-zero score as the system shows some ability to perform translation. However, for GEC, simply copying the input can result in matching several n-grams from the reference despite the system showing zero ability to perform correction. Moreover, the GLEU score will vary with the length of the input sentence without the system having to fix any error or even modify the input. Similarly, for Example 2, Hypothesis 1 and 2 (Table 4), despite being grammatical and matching one of the reference corrections exactly, GLEU gives a non-perfect score for both hypotheses, whereas I and M^2 correctly gives perfect scores. This is an artefact of GLEU randomly selecting one among the two references for scoring and averaging over multiple iterations. Also, due to its non-deterministic behavior, the scores for Hypothesis 1 and 2 differ despite having the same n-gram statistics compared to the references.

Intuitiveness: In Example 1, GLEU produces a non-zero score for a hypothesis that is exactly the same as the input sentence (Hypothesis 1). When an incorrect change is introduced (Hypothesis 2), the score becomes even higher (GLEU=0.735). This is counter-intuitive. If a change results in an ungrammatical sentence, the score should remain the same or decrease as in the case of M^2 and I. This unintuitive behavior of GLEU is due to the additional term in GLEU that penalizes preservation of n-grams in the input sentence that were supposed to be changed according to the reference. It can be argued that GLEU intends to reward GEC systems for detecting errors by assigning a partial credit to systems that make spurious changes at locations where corrections are deemed necessary by human annotators. However, this will encourage building GEC systems that provide inaccurate feedback and potentially mislead the end users (primarily language learners). Hence, it is better to build and evaluate grammatical error detection systems separately (Rei and Yannakoudakis, 2016). I-measure, on the other hand, assigns a negative score for Hypothesis 2 as it is considered to 'degrade' the input, although it is arguable that both hypotheses 1 and 2 are equally ungrammatical. In Example 2, when multiple references are used, GLEU gives a higher score to Hypothesis 3 (ungrammatical) than Hypothesis 1 and 2, both of which are grammatical and each matches one of the two references exactly. On the other hand, both M^2 and I-measure assign a lower score to Hypothesis 3 and conform to our intuition.

5 Related Work

5.1 GEC Evaluation

When GEC was restricted to specific error types, measures such as accuracy, precision, recall, and F-score were employed (Chodorow et al., 2012) as done in the earlier shared tasks (Dale and Kilgarriff, 2011; Dale et al., 2012). Dahlmeier and Ng (2012) identified weaknesses in evaluation methods used in these shared tasks in extracting phrase-level edits that correctly match the reference and proposed MaxMatch scoring (Dahlmeier and Ng, 2012), which can also work for sentence correction over all error types. M^2 does not distinguish between a do-nothing baseline and a system that changes the input incorrectly, and hence I-measure (Felice and Briscoe, 2015) was proposed. Later, GLEU (Napoles et al., 2015) was proposed. It relied on whole sentence rewrites instead of span-based error annotations. However, our analysis shows that GLEU has several weaknesses in its formulation and has a non-deterministic behavior that makes it an unreliable alternative to prior metrics. A few reference-less evaluation methods have been proposed of which (Napoles et al., 2016b) considers the grammaticality of the output sentences alone to evaluate GEC systems. Recently, Asano et al. (2017) proposed improvements to (Napoles et al., 2016b) by additionally accounting for fluency and meaning preservation. Sakaguchi et al. (2017) suggested future directions of improving GEC evaluation such as considering whole document rewrites.

5.2 Human Judgments and Metric Quality

Inspired from WMT, two collections of human judgments of GEC system outputs were released (Grundkiewicz et al., 2015; Napoles et al., 2015). We use the former in our study as it has a much higher number of human pairwise comparisons (109,098) compared to the latter (28,146). Also, (Napoles et al., 2015) includes references as one of the compared systems in order to act as a control measure to ensure quality of human judgments. However, this is unfair to systems that compare more often against the reference as noted by Callison-Burch et al. (2012), since human raters may prefer the reference corrections more often. This bias is further worsened by explicitly displaying the reference corrections to the human judges in (Napoles et al., 2015). Reference translations are neither included in rankings nor shown to judges in (Grundkiewicz et al., 2015). Also, all subsequent studies used the judgments released by Grundkiewicz et al. (2015) for comparing GEC metrics. Grundkiewicz et al. (2015), however, did not compare to GLEU as it was not available at the time. Later work which measured system-level correlation using GLEU failed for a number of reasons which motivate this paper. Apart from using different variations of GLEU, there was a mistake[6] in GLEU computation that produced different results and incorrect conclusions in earlier studies. Moreover, no proper significance tests to compare the differences in correlations between metrics were done. Significance tests that reject the null hypothesis of having no correlation to humans is not adequate for comparing differences in correlations between metrics (Graham and Baldwin, 2014). Also, (Napoles et al., 2016b; Asano et al., 2017) use 18 references for CoNLL-2014 sentences for generating metric rankings, of which 2 are the references used in the shared task, 8 are from (Bryant and Ng, 2015), and another 8 are annotated by both experts and non-experts as sentence-rewrites (Sakaguchi

[6] The brevity penalty was incorrectly implemented as $\exp(1-l_h/l_r)$ instead of $\exp(1-l_r/l_h)$. This was fixed in the version that we use (fixed on 10 June, 2017: https://github.com/cnap/gec-ranking/commit/50b503)

et al., 2016). Automatically constructing M^2 and I-measure annotations from sentence rewrites will be sub-optimal. Also, we observed that the non-determinism of GLEU scores is more pronounced when more number of references are used, resulting in large differences in correlation values. In the end, we decided to use the standard two references for CoNLL-2014 as used in the original shared task and continues to be the standard benchmark used to evaluate GEC systems to date. Also, prior studies had not conducted sentence-level agreement with human judgments, which we did in this paper. Sakaguchi et al. (2017) qualitatively compared metrics using contrived, gamed examples, which are rather irrelevant in practice (such as a system that outputs a dummy hypothesis “a”, “a a”, or “a a a” for any input sentence) and showed that M^2 under-penalized such systems. On the contrary, our qualitative assessment is based on naturally occurring examples and highlights the practical strengths and weaknesses of current GEC metrics.

6 Conclusion

Through carefully designed experiments and significance tests, we find no evidence of GLEU being a better metric than M^2 for ranking systems as claimed in prior work (Napoles et al., 2015; Sakaguchi et al., 2016; Napoles et al., 2016b). In fact, at the sentence level, when correctly predicting human ties are rewarded, M^2 works better than GLEU. We believe that GEC metrics should be interpretable and must provide intuitive scores. In our qualitative assessment, we find that GLEU is less interpretable and produces absurd scores in common scenarios. Our analysis suggests that GLEU cannot be considered as a replacement of existing GEC metrics. The code to replicate the evaluations in this paper is available at `https://github.com/nusnlp/gecmetrics`.

Acknowledgements

We thank the three anonymous reviewers for their useful feedback.

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).*

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation.*

Ondřej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016a. Ten years of WMT evaluation campaigns: Lessons learnt. In *Proceedings of LREC Workshop Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem.*

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016b. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation.*

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation.*

Christopher Bryant and Hwee Tou Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).*

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).*

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation.*

Martin Chodorow, Markus Dickinson, Ross Israel, and Joel Tetreault. 2012. Problems in evaluating grammatical error detection systems. In *Proceedings of the 24th International Conference on Computational Linguistics.*

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*.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*.

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*.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

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*.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

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)*.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2016a. GLEU without tuning. *arXiv preprint arXiv:1605.02592*.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016b. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*.

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*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*.

Marek Rei and Helen Yannakoudakis. 2016. Compositional sequence labeling models for error detection in learner writing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Alla Rozovskaya, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid, and Behrang Mohit. 2015. The second QALB shared task on automatic text correction for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, 4:169–182.

Keisuke Sakaguchi, Courtney Napoles, and Joel Tetreault. 2017. GEC into the future: Where are we going and how do we get there? In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*.

Evan James Williams. 1959. *Regression Analysis*. Wiley.