Ground Truth for Grammatical Error Correction Metrics

Courtney Napoles[1] and **Keisuke Sakaguchi**[1] and **Matt Post**[2] and **Joel Tetreault**[3]
[1]Center for Language and Speech Processing, Johns Hopkins University
[2]Human Language Technology Center of Excellence, Johns Hopkins University
[3]Yahoo Labs

Abstract

How do we know which grammatical error correction (GEC) system is best? A number of metrics have been proposed over the years, each motivated by weaknesses of previous metrics; however, the metrics themselves have not been compared to an empirical gold standard grounded in human judgments. We conducted the first *human evaluation* of GEC system outputs, and show that the rankings produced by metrics such as MaxMatch and I-measure do not correlate well with this ground truth. As a step towards better metrics, we also propose GLEU, a simple variant of BLEU, modified to account for both the source and the reference, and show that it hews much more closely to human judgments.

1 Introduction

Automatic metrics are a critical component for all tasks in natural language processing. For many tasks, such as parsing and part-of-speech tagging, there is a single correct answer, and thus a single metric to compute it. For other tasks, such as machine translation or summarization, there is no effective limit to the size of the set of correct answers. For such tasks, metrics proliferate and compete with each other for the role of the dominant metric. In such cases, an important question to answer is by what means such metrics should be compared. That is, what is the *metric* metric?

The answer is that it should be rooted in the end-use case for the task under consideration. This could be some other metric further downstream of the task, or something simpler like direct human evaluation. This latter approach is the one often taken in machine translation; for example, the organizers of the Workshop on Statistical Machine Translation have long argued that human evaluation is the ultimate ground truth, and have therefore conducted an extensive human evaluation to produce a system ranking, which is then used to compare metrics (Bojar et al., 2014).

Unfortunately, for the subjective task of grammatical error correction (GEC), no such ground truth has ever been established. Instead, the rankings produced by new metrics are justified by their correlation with explicitly-corrected errors in one or more references, and by appeals to intuition for the resulting rankings. However, arguably even more so than for machine translation, the use case for grammatical error correction is human consumption, and therefore, the ground truth ranking should be rooted in human judgments.

We establish a ground truth for GEC by conducting a human evaluation and producing a *human* ranking of the systems entered into the CoNLL-2014 Shared Task on GEC. We find that existing GEC metrics correlate very poorly with the ranking produced by this human evaluation. As a step in the direction of better metrics, we develop the Generalized Language Evaluation Understanding metric (GLEU) inspired by BLEU, which correlates much better with the human ranking than current GEC metrics.[1]

2 Grammatical error correction metrics

GEC is often viewed as a matter of correcting isolated grammatical errors, but is much more complicated, nuanced, and subjective than that. As discussed in Chodorow et al. (2012), there is often no single correction for an error (e.g., whether to correct a subject-verb agreement error by changing the number of the subject or the verb), and errors cover a range of factors including style, register, venue, audience, and usage questions, about

[1] Our code and rankings of the CoNLL-2014 Shared Task system outputs can be downloaded from `github.com/cnap/gec-ranking/`.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 588–593,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

which there can be much disagreement. In addition, errors are not always errors, as can be seen from the existence of different style manuals at newspapers, and questions about the legitimacy of prescriptivist grammar conventions.

Several automatic metrics have been used for evaluating GEC systems. F-score, the harmonic mean of precision and recall, is one of the most commonly used metrics. It was used as an official evaluation metric for several shared tasks (Dale et al., 2012; Dale and Kilgarriff, 2011), where participants were asked to detect and correct closed-class errors (i.e., determiners and prepositions).

One of the issues with F-score is that it fails to capture phrase-level edits. Thus Dahlmeier and Ng (2012) proposed the MaxMatch (M^2) scorer, which calculates the F-score over an edit lattice that captures phrase-level edits. For GEC, M^2 is the standard, having been used to rank error correction systems in the 2013 and 2014 CoNLL shared tasks, where the error types to be corrected were not limited to closed-class errors. (Ng et al., 2013; Ng et al., 2014). M^2 was assessed by comparing its output against that of the official Helping Our Own (HOO) scorer (Dale and Kilgarriff, 2011), itself based on the GNU `wdiff` utility.[2] In other words, it was evaluated under the assumption that evaluating GEC can be reduced to checking whether a set of predefined errors have been changed into a set of associated corrections.

M^2 is not without its own issues. First, phrase-level edits can be gamed because the lattice treats a long phrase deletion as one edit.[3] Second, the F-score does not capture the difference between "no change" and "wrong edits" made by systems. Chodorow et al. (2012) also list other complications arising from using F-score or M^2, depending on the application of GEC.

Considering these problems, Felice and Briscoe (2015) proposed a new metric, I-measure, which is based on accuracy computed by edit distance between the source, reference, and system output. Their results are striking: there is a negative correlation between the M^2 and I-measure scores (Pearson's $r = -0.694$).

A difficulty with all these metrics is that they require detailed annotations of the location and error type of each correction in response to an explicit error annotation scheme. Due to the inherent subjectivity and poor definition of the task, mentioned above, it is difficult for annotators to reliably produce these annotations (Bryant and Ng, 2015). However, this requirement can be relinquished by treating GEC as a text-to-text rewriting task and borrowing metrics from machine translation, as Park and Levy (2011) did with BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007).

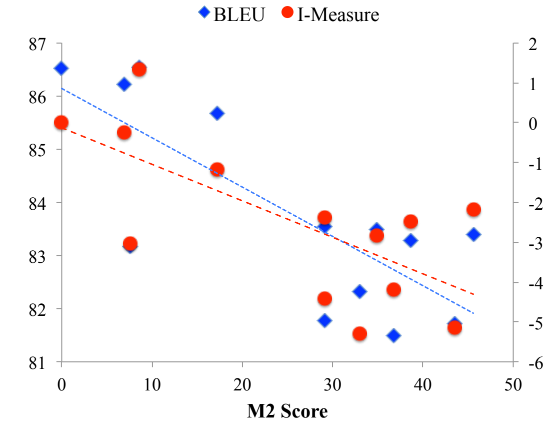


Figure 1: Correlation among M^2, I-measure, and BLEU scores: M^2 score shows negative correlations to other metrics.

As we will show in more detail in Section 5, taking the twelve publicly released system outputs from the CoNLL-2014 Shared Task,[4] we actually find a negative correlation between the M^2 and BLEU scores ($r = -0.772$) and positive correlation between I-measure and BLEU scores ($r = 0.949$) (Figure 1). With the earlier-reported negative correlation between I-measure and M^2, we have a troubling picture: which of these metrics is best? Which one actually captures and rewards the behaviors we would like our systems to report? Despite these many proposed metrics, no prior work has attempted to answer these questions by comparing them to human judgments. We propose to answer these questions by producing a definitive human ranking, against which the rankings of different metrics can be compared.

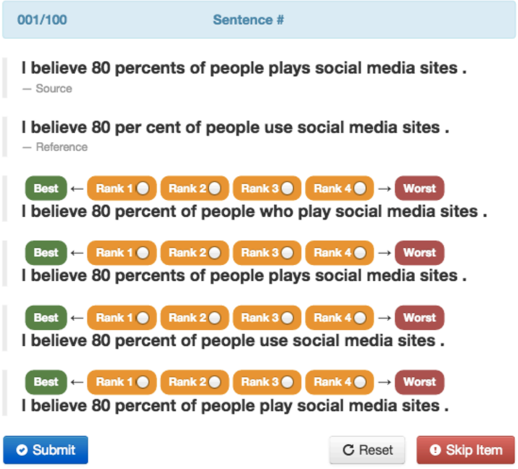
3 The human ranking

The Workshop on Statistical Machine Translation (WMT) faces the same question each year as part

[2] `http://www.gnu.org/s/wdiff/`

[3] For example, when we put a single character 'X' as system output for each sentence, we obtain $P = 0.27, R = 0.29, M^2 = 0.28$, which would be ranked 6/13 systems in the 2014 CoNLL shared task.

[4] `www.comp.nus.edu.sg/~nlp/conll14st.html`



I believe 80 percents of people plays social media sites .
— Source

I believe 80 per cent of people use social media sites .
— Reference

I believe 80 percent of people who play social media sites .

I believe 80 percents of people plays social media sites .

I believe 80 percent of people use social media sites .

I believe 80 percent of people play social media sites .

Figure 2: The Appraise evaluation system.

of its metrics shared task. Arguing that humans are the ultimate judge of quality, they gather human judgments and use them to produce a ranking of the systems for each task. Machine translation metrics are then evaluated based on how closely they match this ranking, using Pearson's r (prior to 2014) or Spearman's ρ (2014).

We borrow their approach to conduct a human evaluation. We used Appraise (Federmann, 2012)[5] to collect pairwise judgments among 14 systems: the output of 12 systems entered in the CoNLL-14 Shared Task, plus the source and a reference sentence. Appraise presents the judge with the source and reference sentence[6] and asks her to rank four randomly selected systems from best to worst, ties allowed (Figure 2). The four-way ranking is transformed into a set of pairwise judgments.

We collected data from three native English speakers, resulting in 28,146 pairwise system judgements. Each system's quality was estimated and the total ranking was produced on this dataset using the TrueSkill model (Sakaguchi et al., 2014), as done in WMT 2014. The annotators had strong correlations in terms of the total system ranking and estimated quality, with the reference being ranked at the top (Table 1).

Judges	r	ρ
1 and 2	0.80	0.69
1 and 3	0.73	0.80
2 and 3	0.81	0.71

Table 1: Pearson's r and Spearman's ρ correlations among judges (excluding the reference).

4 Generalized BLEU

Current metrics for GEC rely on references with explicitly labeled error annotations, the type and form of which vary from task to task and can be difficult to convert. Recognizing the inherent ambiguity in the error-correction task, a better metric might be independent of such an annotation scheme and only require corrected references. This is the view of GEC as a generic text-rewriting task, and it is natural to apply standard metrics from machine translation. However, applied off-the-shelf, these metrics yield unintuitive results. For example, BLEU ranks the *source* sentence as second place in the CoNLL-2014 shared task.[7]

The problem is partially due to the subtle but important difference between machine translation and monolingual text-rewriting tasks. In MT, an untranslated word or phrase is almost always an error, but in grammatical error correction, this is not the case. Some, but not all, regions of the source sentence should be changed. This observation motivates a small change to BLEU that computes n-gram precisions over the reference but assigns more weight to n-grams that have been correctly changed from the source. This revised metric, Generalized Language Evaluation Understanding (GLEU), rewards corrections while also correctly crediting unchanged source text.

Recall that BLEU(C, R) (Papineni et al., 2002) is computed as the geometric mean of the modified precision scores of the test sentences C relative to the references R, multiplied by a brevity penalty to control for recall. The precisions are computed over bags of n-grams derived from the candidate translation and the references. Each n-gram in the candidate sentence is "clipped" to the maximum count of that n-gram in any of the references, ensuring that no precision is greater than 1.

Similar to I-measure, which calculates a weighted accuracy of edits, we calculate a weighted precision of n-grams. In our adaptation, we modify the precision calculation to assign extra weight to n-grams present in the candidate that overlap with the reference *but not* the source (the set of n-grams $R \setminus S$). The precision is also penal-

[5] `github.com/cfedermann/Appraise`

[6] CoNLL-14 has two references. For each sentence, we randomly chose one to present as the answer and one to be among the systems to be ranked.

[7] Of course, it could be the case that the source sentence is actually the second best, but our human evaluation (§5) confirms that this is not the case.

$$p'_n = \frac{\sum\limits_{n\text{-}gram \in C} Count_{R\setminus S}(n\text{-}gram) - \lambda\left(Count_{S\setminus R}(n\text{-}gram)\right) + Count_R(n\text{-}gram)}{\sum\limits_{n\text{-}gram' \in C'} Count_S(n\text{-}gram') + \sum\limits_{n\text{-}gram \in R\setminus S} Count_{R\setminus S}(n\text{-}gram)} \tag{1}$$

ized by a weighted count of n-grams in the candidate that are in the source but not the reference (false negatives, $S \setminus R$). For a correction candidate C with a corresponding source S and reference R, the modified n-gram precision for GLEU(C,R,S) is shown in Equation 1. The weight λ determines by how much incorrectly changed n-grams are penalized. Equations 2–3 describe how the counts are collected given a bag of n-grams B.

$$Count_B(n\text{-}gram) = \sum_{n\text{-}gram' \in B} d(n\text{-}gram, n\text{-}gram') \tag{2}$$

$$d(n\text{-}gram, n\text{-}gram') = \begin{cases} 1 & \text{if } n\text{-}gram = n\text{-}gram' \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-c/r)} & \text{if } c \leq r \end{cases} \tag{4}$$

$$\text{GLEU}\,(C, R, S) = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p'_n\right) \tag{5}$$

In our experiments, we used $N = 4$ and $w_n = \frac{1}{N}$, which are standard parameters for MT, the same brevity penalty as BLEU (Equation 4), and report results on $\lambda = \{0.1, 0\}$ (GLEU$_{0.1}$ and GLEU$_0$, respectively). For this task, not penalizing false negatives correlates best with human judgments, but the weight can be tuned for different tasks and datasets. GLEU can be easily extended to additionally punish false positives (incorrectly editing grammatical text) as well.

5 Results

The respective system rankings of each metric are presented in Table 2. The human ranking is considerably different from those of most of the metrics, a fact that is also captured in correlation coefficients (Table 3).[8] From the human evaluation, we learn that the source falls near the middle of the rankings, even though the BLEU, I-measure and M^2 rank it among the best or worst systems.

M^2, the metric that has been used for the CoNLL shared tasks, only correlates moderately with human rankings, suggesting that it is not an ideal metric for judging the results of a competition. Even though I-measure perceptively aims to predict whether an output is better or worse than the input, it actually has a slight negative correlation with human rankings. GLEU$_0$ is the only metric that strongly correlates with the human ranks, and performs closest to the range of human-to-human correlation ($0.73 \leq r \leq 0.81$) GLEU$_0$ correctly ranks four out of five of the top human-ranked systems at the top of its list, while the other metrics rank at most three of these systems in the top five.

Metric	r	ρ
GLEU$_0$	**0.542**	**0.555**
M^2	0.358	0.429
GLEU$_{0.1}$	0.200	0.412
I-measure	-0.051	-0.005
BLEU	-0.125	-0.225

Table 3: Correlation of metrics with the human ranking (excluding the reference), as calculated with Pearson's r and Spearman's ρ.

All metrics deviate from the human rankings, which may in part be because automatic metrics equally weight all error types, when some errors may be more tolerable to human judges than others. For example, inserting a missing token is rewarded the same by automatic metrics, whether it is a comma or a verb, while a human would much more strongly prefer the insertion of the latter. An example of system outputs with their automatic scores and human rankings is included in Table 4.

This example illustrates some challenges faced when using automatic metrics to evaluate GEC. The automatic metrics weight all corrections equally and are limited to the gold-standard references provided. Both automatic metrics, M^2 and GLEU, prefer the AMU output in this example, even though it corrects one error and *introduces* another. The human judges rank the UMC output as the best for correcting the main verb even though it ignored the spelling error. The UMC and NTHU sentences both receive M^2 $= 0$ because they make none of the gold-standard edits, even though UMC correctly inserts *be* into the sentence. M^2 does not recognize this since it is in a different location from where the annotators placed it.

[8] Pearson's measure assumes the scores are normally distributed, which may not be true here.

Human	**BLEU**	**I-measure**	**M^2**	**GLEU$_0$**	**GLEU$_{0.1}$**
CAMB	UFC	UFC	CUUI	CUUI	CUUI
AMU	source	source	CAMB	AMU	AMU
RAC	IITB	IITB	AMU	UFC	CAMB
CUUI	SJTU	SJTU	POST	CAMB	UFC
source	UMC	CUUI	UMC	source	IITB
POST	CUUI	PKU	NTHU	IITB	SJTU
UFC	PKU	AMU	PKU	SJTU	PKU
SJTU	AMU	UMC	RAC	PKU	UMC
IITB	IPN	IPN	SJTU	UMC	NTHU
PKU	NTHU	POST	UFC	NTHU	POST
UMC	CAMB	RAC	IPN	POST	RAC
NTHU	RAC	CAMB	IITB	RAC	IPN
IPN	POST	NTHU	source	IPN	source

Table 2: System outputs scored by different metrics, ranked best to worst.

System	**Sentence**	**Scores**
Original sentence	We may in actual fact communicating with a hoax Facebook acccount of a cyber friend , which we assume to be real but in reality , it is a fake account .	–
Reference 1	We may in actual fact **be** communicating with a hoax Facebook acccount of a cyber friend , which we assume to be real but in reality , it is a fake account .	–
Reference 2	We may in actual fact **be** communicating with a **fake** Facebook **account** of **an online** friend , which we assume to be real but , in reality , it is a fake account .	–
UMC	We may **be** in actual fact communicating with a hoax Facebook acccount of a cyber friend , we assume to be real but in reality , it is a fake account .	GLEU = 0.62 M^2 = 0.00 Human rank= 1
AMU	We may in actual fact communicating with a hoax Facebook **account** of a cyber friend , which we assume to be real but in reality , it is a fake **accounts** .	GLEU = 0.64 M^2 = 0.39 Human rank= 2
NTHU	We may of actual fact communicating with a hoax Facebook acccount of a cyber friend , which we **assumed** to be real but in reality , it is a fake account .	GLEU = 0.60 M^2 = 0.00 Human rank= 4

Table 4: Examples of system output (changes are in bold) and the sentence-level scores assigned by different metrics.

However, GLEU awards UMC partial credit for adding the correct unigram, and further assigns all sentences a real score.

6 Summary

As with other metrics used in natural language processing tasks, grammatical error correction metrics must be evaluated against ground truth. The inherent subjectivity in what constitutes a grammatical correction, together with the fact that the use case for grammatically-corrected output is human readers, argue for grounding metric evaluations in a human evaluation, which we produced following procedures established by the Workshop on Statistical Machine Translation. This human ranking shows us that the metric commonly used for GEC is not appropriate, since it does not correlate strongly; newly proposed alternatives fare little better.

Attending to how humans perceive the quality of the sentences, we developed GLEU by making a simple variation to an existing metric. GLEU more closely models human judgments than previous metrics because it rewards correct edits while penalizing ungrammatical edits, while capturing fluency and grammatical constraints by virtue of using n-grams. While this simple modification to BLEU accounts for crucial differences in a monolingual setting, fares well, and could take the place of existing metrics, especially for rapid system development as in machine translation, there is still room for further work as there is a gap in how well it correlates with human judgments.

Most importantly, the results and data from this paper establish a method for objectively evaluating future metric proposals, which is crucial to yearly incremental improvements to the GEC task.

Acknowledgments

We would like to thank Christopher Bryant, Hwee Tou Ng, Mariano Felice, and Ted Briscoe for sharing their research with us pre-publication. We also thank the reviewers and Wei Xu for their valuable feedback, and Benjamin Van Durme for his support.

References

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

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*, Beijing, China, July. Association for Computational Linguistics.

Martin Chodorow, Markus Dickinson, Ross Israel, and Joel Tetreault. 2012. Problems in evaluating grammatical error detection systems. In *Proceedings of COLING 2012*, pages 611–628, Mumbai, India, December. The COLING 2012 Organizing Committee.

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. Association for Computational Linguistics, June.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 242–249, Nancy, France, September. Association for Computational Linguistics.

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*, pages 54–62, Montréal, Canada, June. Association for Computational Linguistics.

Christian Federmann. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35, September.

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*, Denver, CO, June. Association for Computational Linguistics.

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

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*, pages 1–12, Sofia, Bulgaria, August. 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, Baltimore, Maryland, June. Association for Computational Linguistics.

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*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Y. Albert Park and Roger Levy. 2011. Automated whole sentence grammar correction using a noisy channel model. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 934–944, Portland, Oregon, USA, June. Association for Computational Linguistics.

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*, pages 1–11, Baltimore, Maryland, USA, June. Association for Computational Linguistics.