There's No Comparison: Reference-less Evaluation Metrics in Grammatical Error Correction

Courtney Napoles,[1] **Keisuke Sakaguchi**,[1] and **Joel Tetreault**[2]
[1]Center for Language and Speech Processing, Johns Hopkins University
[2]Grammarly
{napoles,keisuke}@cs.jhu.edu, joel.tetreault@grammarly.com

Abstract

Current methods for automatically evaluating grammatical error correction (GEC) systems rely on gold-standard references. However, these methods suffer from penalizing grammatical edits that are correct but not in the gold standard. We show that reference-less grammaticality metrics correlate very strongly with human judgments and are competitive with the leading reference-based evaluation metrics. By interpolating both methods, we achieve state-of-the-art correlation with human judgments. Finally, we show that GEC metrics are much more reliable when they are calculated at the sentence level instead of the corpus level. We have set up a CodaLab site for benchmarking GEC output using a common dataset and different evaluation metrics.

1 Introduction

Grammatical error correction (GEC) has been evaluated by comparing the changes made by a system to the corrections made in gold-standard annotations. Following the recent shared tasks in this field (e.g., Ng et al. (2014)), several papers have critiqued GEC metrics and proposed new methods. Existing metrics depend on gold-standard corrections and therefore have a notable weakness: systems are penalized for making corrections that do not appear in the references.[1] For example, the following output has low metric scores even though three appropriate corrections were made to the input:

> However , people now can ~~contact~~ **communicate** with ~~anyone~~ **people** all over the world who can use computer**s** at any time , and there is no time delay of messages .

These changes (in red) were not seen in the references and therefore the metrics GLEU and M^2 (described in §2) score this output worse than 75% of 15,000 other generated sentences.

While grammaticality-based, reference-less metrics have been effective in estimating the quality of machine translation (MT) output, the utility of such metrics has not been investigated previously for GEC. We hypothesize that such methods can overcome this weakness in reference-based GEC metrics. This paper has four contributions: 1) We develop three *grammaticality* metrics that are competitive with current reference-based measures and correlate very strongly with human judgments. 2) We achieve state-of-the-art performance when interpolating a leading reference-based metric with a grammaticality metric. 3) We identify an interesting result that the mean of sentence-level scores is substantially better for evaluating systems than the system-level score. 4) We release code for two grammaticality metrics and establish an online platform for evaluating GEC output.

2 Prior work

To our knowledge, this is the first work to evaluate GEC without references. Within MT, this task is called *quality estimation*. MT output is evaluated by its *fluency*, or adherence to accepted conventions of grammaticality and style, and *adequacy*, which is the input's meaning conveyed in the output.

[1]We refer to the gold-standard corrections as *references* because *gold standard* suggests just one accurate correction.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2109–2115, Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

Quality estimation targets fluency by estimating the amount of post-editing needed by the output. This has been the topic of recent shared tasks, e.g. Bojar et al. (2015). Specia et al. (2010) evaluated the quality of translations using sentence-level features from the output but not the references, predicting discrete and continuous scores. A strong baseline, *QuEst*, uses support vector regression trained over 17 features extracted from the output (Specia et al., 2013). Most closely related to our work, Parton et al. (2011) applied features from Educational Testing Service's e-rater® (Attali and Burstein, 2006) to evaluate MT output with a ranking SVM, without references, and improved performance by including features derived from MT metrics (BLEU, TERp, and METEOR).

Within the GEC field, recent shared tasks have prompted the development and scrutiny of new metrics for evaluating GEC systems. The Helping Our Own shared tasks evaluated systems using precision, recall, and F-score against annotated gold-standard corrections (Dale and Kilgarriff, 2011; Dale et al., 2012). The subsequent CoNLL Shared Tasks on GEC (Ng et al., 2013; Ng et al., 2014) were scored with the MaxMatch metric (M^2), which captures word- and phrase-level edits by calculating the F-score over an edit lattice (Dahlmeier and Ng, 2012). Felice and Briscoe (2015) identified shortcomings of M^2 and proposed I-measure to address these issues. I-measure computes the accuracy of a token-level alignment between the original, generated, and gold-standard sentences. These precision- and recall-based metrics measure fluency and adequacy by penalizing inappropriate changes, which alter meaning or introduce other errors. Changes consistent with the annotations indicate improved fluency and no change in meaning.

Unlike these metrics, GLEU scores output by penalizing n-grams found in the input and output but not the reference (Napoles et al., 2015). Like BLEU (Papineni et al., 2002), GLEU captures both fluency and adequacy with n-gram overlap. Recent work has shown that GLEU has the strongest correlation with human judgments compared to the GEC metrics described above (Sakaguchi et al., 2016). These GEC metrics are all defined at the corpus level, meaning that the statistics are accumulated over the entire output and then used to calculate a single system score.

3 Explicitly evaluating grammaticality

GLEU, I-measure, and M^2 are calculated based on comparison to reference corrections. These Reference-Based Metrics (*RBMs*) credit corrections seen in the references and penalize systems for ignoring errors and making bad changes (changing a span of text in an ungrammatical way or introducing errors to grammatical text). However, RBMs make two strong assumptions: that the annotations in the references are *correct* and that they are *complete*. We argue that these assumptions are invalid and point to a deficit in current evaluation practices. In GEC, the agreement between raters can be low due to the challenging nature of the task (Bryant and Ng, 2015; Rozovskaya and Roth, 2010; Tetreault and Chodorow, 2008), indicating that annotations may not be correct or complete.

An exhaustive list of all possible corrections would be time-consuming, if not impossible. As a result, RBMs penalize output that has a valid correction that is not present in the references or that addresses an error not corrected in the references. The example in §1 has low GLEU and M^2 scores, even though the output addresses two errors (GLEU=0.43 and $M^2 = 0.00$, in the bottom half and quartile of 15k system outputs, respectively).

To address these concerns, we propose three metrics to evaluate the grammaticality of output without comparing to the input or a gold-standard sentence (Grammaticality-Based Metrics, or *GBMs*). We expect GBMs to score sentences, such as our example in §1, more highly. The first two metrics are scored by counting the errors found by existing grammatical error detection tools. The error count score is simply calculated: $1 - \frac{\#\text{ errors}}{\#\text{ tokens}}$. Two different tools are used to count errors: e-rater®'s grammatical error detection modules (ER) and Language Tool (Miłkowski, 2010) (LT). We choose these because, while e-rater® is a large-scale, robust tool that detects more errors than Language Tool,[2] it is proprietary whereas Language Tool is publicly available and open sourced.

For our third method, we estimate a grammaticality score with a linguistic feature-based model (LFM), which is our implementation of Heilman et

[2]In the data used for this work, e-rater® detects approximately 15 times more errors than Language Tool.

al. (2014).[3] The LFM score is a ridge regression over a variety of linguistic features related to grammaticality, including the number of misspellings, language model scores, OOV counts, and PCFG and link grammar features. It has been shown to effectively assess the grammaticality of learner writing. LFM predicts a score for each sentence while ER and LT, like the RBMs, can be calculated with either sentence- or document-level statistics. To be consistent with LFM, for all metrics in this work we score each sentence individually and report the system score as the mean of the sentence scores. We discuss the effects of modifying metrics from a corpus-level to a sentence-level in §5.

Consistent with our hypothesis, ER and LT score the §1 example in the top quartile of outputs and LFM ranks it in the top half.

3.1 A hybrid metric

The obvious deficit of GBMs is that they do not measure the adequacy of generated sentences, so they could easily be manipulated with grammatical output that is unrelated to the input. An ideal GEC metric would measure both the grammaticality of a generated sentence and its meaning compared to the original sentence, and would not necessarily need references. The available data of scored system outputs are insufficient for developing a new metric due to their limited size, thus we turn to interpolation to develop a sophisticated metric that jointly captures grammaticality and adequacy.

To harness the advantage of RBMs (adequacy) and GBMs (fluency), we build combined metrics, interpolating each RBM with each GBM. For a sentence of system output, the interpolated score (S_I) of the GBM score (S_G) and RBM score (S_R) is computed as follows:

$$S_\mathrm{I} = (1 - \lambda)S_\mathrm{G} + \lambda S_\mathrm{R}$$

All values of S_G and S_R are in the interval $[0, 1]$, except for I-measure, which falls between $[-1, 1]$, and the distribution varies for each metric.[4] The system score is the average S_I of all generated sentences.

[3] Our implementation is slightly modified in that it does not use features from the PET HPSG parser.

[4] Mean scores are GLEU 0.52 ± 0.21, M^2 0.21 ± 0.34, IM 0.10 ± 0.30, ER 0.91 ± 0.10, LFM 0.50 ± 0.16, LT 1.00 ± 0.01.

Metric	Spearman's ρ	Pearson's r
GLEU	**0.852**	**0.838**
ER	**0.852**	0.829
LT	0.808	0.811
I-measure	0.769	0.753
LFM	0.780	0.742
M^2	0.648	0.641

Table 1: Correlation between the human and metric rankings.

4 Experiments

To assess the proposed metrics, we apply the RBMs, GBMs, and interpolated metrics to score the output of 12 systems participating in the CoNLL-2014 Shared Task on GEC (Ng et al., 2014). Recent works have evaluated RBMs by collecting human rankings of these system outputs and comparing them to the metric rankings (Grundkiewicz et al., 2015; Napoles et al., 2015). In this section, we compare each metric's ranking to the human ranking of Grundkiewicz et al. (2015, Table 3c). We use 20 references for scoring with RBMs: 2 original references, 10 references collected by Bryant and Ng (2015), and 8 references collected by Sakaguchi et al. (2016). The motivations for using 20 references are twofold: the best GEC evaluation method uses these 20 references with the GLEU metric (Sakaguchi et al., 2016), and work in machine translation shows that more references are better for evaluation (Finch et al., 2004). Due to the low agreement discussed in §3, having more references can be beneficial for evaluating a system when there are multiple viable ways of correcting a sentence. Unlike previous GEC evaluations, all metrics reported here use the *mean* of the sentence-level scores for each system.

Results are presented in Table 1. The error-count metrics, ER and LT, have stronger correlation than all RBMs except for GLEU, which is the current state of the art. GLEU has the strongest correlation with the human ranking ($\rho = 0.852$, $r = 0.838$), followed closely by ER, which has slightly lower Pearson correlation ($r = 0.829$) but equally as strong rank correlation, which is arguably more important when comparing different systems. I-measure and LFM have similar strength correlations, and M^2 is the lowest performing metric, even though its correlation is still strong ($\rho = 0.648$, $r = 0.641$).

Next we compare the interpolated scores with the human ranking, testing 101 different values of λ

	ranked by Spearman's rank coefficient (ρ)				*ranked by Pearson's correlation coefficient (r)*			
		ER	**LFM**	**LT**		**ER**	**LFM**	**LT**
	no intrpl.	0.852 (0)	0.780 (0)	0.808 (0)	*no intrpl.*	0.829 (0)	0.742 (0)	0.811 (0)
GLEU	0.852 (1)	**0.885** (0.03)	0.874 (0.27)	0.857 (0.04)	0.838 (1)	**0.867** (0.27)	0.845 (0.84)	0.867 (0.09)
I-m.	0.769 (1)	0.874 (0.19)	0.863 (0.37)	0.852 (0.01)	0.753 (1)	0.837 (0.02)	0.791 (0.22)	0.828 (0.01)
M^2	0.648 (1)	0.868 (0.01)	0.852 (0.05)	0.808 (0.00)	0.641 (1)	0.829 (0.00)	0.754 (0.04)	0.811 (0.00)

Table 2: Oracle correlations between interpolated metrics and the human rankings. The λ value for each metric is in parentheses.

GLEU rank	**Intrpl. rank**	**Output sentence**
1	2	*Genectic* testing is a personal decision , with the knowledge that there is a *possiblity* that one could be a carrier or not .
2	3	*Genectic* testing is a personal decision , the *kowledge* that there is a *possiblity* that one could be a carrier or not .
3	1	Genetic testing is a personal decision , with the knowledge that there is a possibility that one could be a carrier or not .

Table 3: An example of system outputs ranked by GLEU and GLEU interpolated with ER. Words in italics are misspelled.

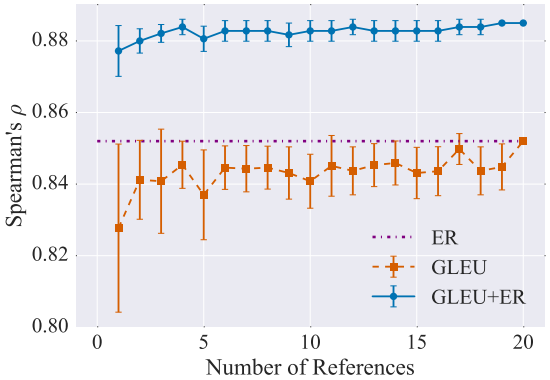

Figure 1: The mean correlation of oracle interpolated GLEU and ER scores across different reference sizes, compared to the uninterpolated metrics. Bars indicate a 95% confidence interval.

[0,1] to find the oracle value. Table 2 shows the correlations between the human judgments and the oracle interpolated metrics. Correlations of interpolated metrics are the upper bound and the correlations of uninterpolated metrics (in the first column and first row) are the lower bound. Interpolating GLEU and IM with GBMs has stronger correlation than any uninterpolated metric (i.e. $\lambda = 0$ or 1), and the oracle interpolation of ER and GLEU manifests the strongest correlation with the human ranking ($\rho = 0.885$, $r = 0.867$).[5] M^2 has the weakest correlation of all uninterpolated metrics and, when combined with GBMs, three of the interpolated metrics have $\lambda = 0$, meaning the interpolated score is equivalent to the GBM and M^2 does not contribute.

Table 3 presents an example of how interpolation can help evaluation. The top two sentences ranked by GLEU have misspellings that were not corrected in the NUCLE references. Interpolating with a GBM rightly ranks the misspelled output below the corrected output.

Since these experiments use a large number of references (20), we determine how different reference sizes affect the interpolated metrics by systematically increasing the number of references from 1 to 20 and randomly choosing n references to use as a gold standard when calculating the RBM scores, repeating 10 times for each value n (Figure 1). The correlation is nearly as strong with 3 and 20 references ($\rho = 0.884$ v. 0.885), and interpolating with just 1 reference is nearly as good (0.878) and improves over any uninterpolated metric.

We acknowledge that using GBMs is in effect using GEC systems to score other GEC systems. Interestingly, we find that even if the GBMs are imperfect, they still boost performance of the RBMs. GBMs have been trained to recognize errors in different contexts and, conversely, can identify correct grammatical constructions in diverse contexts, where the RBMs only recognize corrections made that appear in the gold standards, which are limited. Therefore GBMs can make a nice complement to shortcomings that RBMs may have.

5 Sentence-level evaluation

In the course of our experiments, we noticed that I-measure and GLEU have stronger correlations with the expert human ranking when using the

[5] To verify that these metrics cannot be gamed, we interpolated GBMs with RBMs scored against randomized references, and got scores 15% lower than un-gamed scores, on average.

	Corpus		Sentence	
Metric	ρ	r	ρ	r
GLEU	0.725	0.724	0.852	0.838
I-m.	-0.055*	0.061	0.769	0.753
M^2	0.692	0.617	0.648	0.641

Table 4: Correlation with human ranking when using corpus- and sentence-level metrics. * indicates a significant difference from the corresponding sentence-level correlation ($p < 0.05$).[7]

mean sentence-level score (Table 4).[6] Most notably, I-measure does not correlate at all as a corpus-level metric but has a very strong correlation at the sentence-level (specifically, ρ improves from -0.055 to 0.769). This could be because corpus-level statistics equally distribute counts of correct annotations over all sentences, even those where the output neglects to make any necessary corrections. Sentence-level statistics would not average the correct counts over all sentences in this same way. As a result, a corpus-level statistic may over-estimate the quality of system output. Deeper investigation into this phenomenon is needed to understand why the mean sentence-level scores do better.

6 Summary

We have identified a shortcoming of reference-based metrics: they penalize changes made that do not appear in the references, even if those changes are acceptable. To address this problem, we developed metrics to explicitly measure grammaticality without relying on reference corrections and showed that the error-count metrics are competitive with the best reference-based metric. Furthermore, by interpolating RBMs with GBMs, the system ranking has even stronger correlation with the human ranking. The ER metric, which was derived from counts of errors detected using e-rater®, is nearly as good as the state-of-the-art RBM (GLEU) and the interpolation of these metrics has the strongest reported correlation with the human ranking ($\rho = 0.885$, $r = 0.867$). However, since e-rater® is not widely available to researchers, we also tested a metric using Language Tool, which does nearly as well when interpolated with GLEU ($\rho = 0.857$, $r = 0.867$).

Two important points should be noted: First, due to the small sample size (12 system outputs), none of the rankings significantly differ from one another except for the corpus-level I-measure. Secondly, GLEU and the other RBMs already have strong to very strong correlation with the human judgments, which makes it difficult for any combination of metrics to perform substantially higher. The best uninterpolated and interpolated metrics use an extremely large number of references (20), however Figure 1 shows that interpolating GLEU using just one reference has stronger correlation than any uninterpolated metric. This supports the use of interpolation to improve GEC evaluation in any setting.

This work is the first exploration into applying fluency-based metrics to GEC evaluation. We believe that, for future work, fluency measures could be further improved with other methods, such as using existing GEC systems to *detect* errors, or even using an ensemble of systems to improve coverage (indeed, ensembles have been useful in the GEC task itself (Susanto et al., 2014)). There is also recent work from the MT community, such as the use of confidence bounds (Graham and Liu, 2016) or uncertainty measurement (Beck et al., 2016), which could be adopted by the GEC community.

Finally, in the course of our experiments, we determined that metrics calculated on the sentence-level is more reliable for evaluating GEC output, and we suggest that the GEC community adopt this modification to better assess systems.

To facilitate GEC evaluation, we have set up an online platform[8] for benchmarking system output on the same set of sentences evaluated using different metrics and made the code for calculating LT and LFM available.[9]

Acknowledgments

We would like to thank Matt Post, Martin Chodorow, and the three anonymous reviews for their comments and feedback. We also thank Educational Testing Service for providing e-rater® output. This material is based upon work partially supported by the NSF GRF under Grant No. 1232825.

[6] The correlations in Table 4 differ from what was reported in Grundkiewicz et al. (2015) and Napoles et al. (2015) due to the references and sentence-level computation used in this work.

[7] Significance is found by applying a two-tailed t-test to the Z-scores attained using Fisher's z-transformation.

[8] https://competitions.codalab.org/competitions/12731

[9] https://github.com/cnap/grammaticality-metrics

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Daniel Beck, Lucia Specia, and Trevor Cohn. 2016. Exploring prediction uncertainty in machine translation quality estimation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 208–218, Berlin, Germany, August. Association for Computational Linguistics.

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*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ondřej 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.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. 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 and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 697–707, Beijing, China, July. Association for Computational Linguistics.

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*, pages 10–51, Montréal, Canada, June. Association for 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*, pages 568–572, Montréal, Canada, June. Association for Computational Linguistics.

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.

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*, pages 578–587, Denver, Colorado, June. Association for Computational Linguistics.

Andrew M. Finch, Yasuhiro Akiba, and Eiichiro Sumita. 2004. How does automatic machine translation evaluation correlate with human scoring as the number of reference translations increases? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*.

Yvette Graham and Qun Liu. 2016. Achieving accurate conclusions in evaluation of automatic machine translation metrics. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–10, San Diego, California, June. Association for Computational Linguistics.

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, Lisbon, Portugal, September. Association for Computational Linguistics.

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, Baltimore, Maryland, June. Association for Computational Linguistics.

Marcin Miłkowski. 2010. Developing an open-source, rule-based proofreading tool. *Software: Practice and Experience*, 40(7):543–566.

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, Beijing, China, July. 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.

Kristen Parton, Joel Tetreault, Nitin Madnani, and Martin Chodorow. 2011. E-rating machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 108–115. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2010. Annotating ESL errors: Challenges and rewards. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36, Los Angeles, California, June. Association for Computational Linguistics.

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.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst – a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria, August. Association for Computational Linguistics.

Raymond Hendy Susanto, Peter Phandi, and Hwee Tou Ng. 2014. System combination for grammatical error correction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 951–962, Doha, Qatar, October. Association for Computational Linguistics.

Joel Tetreault and Martin Chodorow. 2008. Native judgments of non-native usage: Experiments in preposition error detection. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 24–32, Manchester, UK, August. Coling 2008 Organizing Committee.