Enabling Robust Grammatical Error Correction in New Domains: Data Sets, Metrics, and Analyses

Courtney Napoles and **Maria Nădejde** and **Joel Tetreault**
Grammarly
Courtney.Napoles@grammarly.com

Abstract

Until now, grammatical error correction (GEC) has been primarily evaluated on text written by non-native English speakers, with a focus on student essays. This paper enables GEC development on text written by native speakers by providing a new data set and metric. We present a multiple-reference test corpus for GEC that includes 4,000 sentences in two new domains (*formal* and *informal* writing by native English speakers) and 2,000 sentences from a diverse set of non-native *student writing*. We also collect human judgments of several GEC systems on this new test set and perform a meta-evaluation, assessing how reliable automatic metrics are across these domains. We find that commonly used GEC metrics have inconsistent performance across domains, and therefore we propose a new ensemble metric that is robust on all three domains of text.

1 Introduction

Grammatical error correction (GEC) systems are evaluated with automatic metrics that compare their output to gold-standard corrections from reference corpora. Having automatic metrics that correlate well with human judgments allows rapid system development and reliable evaluations across the field. Although the GEC community has benefited from several evaluation sets over the past several years, they are primarily composed of student essays written by non-native English speakers (Yannakoudakis et al., 2011; Dahlmeier et al., 2013; Napoles et al., 2017; Bryant et al., 2019). As of yet, we do not know how well GEC systems do on other domains of text or how reliable automatic evaluation is when we move to other domains.

We tackle this issue head-on by creating a new test set for GEC that represents diverse domains of text. We call this new test set *Grammarly Multi-domain Evaluation for GEC Data set* (GMEG-Data), and it is the first test set to include multiple corrections of sentences written by native English speakers and informal writing. We collect human ratings of the corrections generated by 6 GEC systems on this data. From this gold-standard human evaluation, we measure the performance of current GEC metrics and find that the standard metrics are not robust across domains. We propose *Grammarly Multi-domain Evaluation for GEC Metric* (GMEG-Metric), a new ensemble scorer that is reliable on all three domains and draws from the statistics of existing metrics.

This work takes the first major step to develop truly robust GEC systems. Our contributions include:

- GMEG-Data: The largest multiple-reference GEC test set with 6,000 sentences from three domains: *formal* and *informal* text by native English speakers, and a diverse set of *student writing*.
- The output of 6 GEC systems with varying modern architectures and training data sizes, and human judgments of the entire set of system outputs.
- Evaluation of 4 standard GEC metrics and a leading machine translation (MT) metric on the new test set.
- GMEG-Metric: A new ensemble metric for GEC that is robust across all three domains.
- GMEG-Data, GMEG-Metric, human ratings, and system outputs will be made public to enable, for the first time, further development of domain-robust GEC metrics and systems.[1]

[1] https://github.com/grammarly/GMEG

Transactions of the Association for Computational Linguistics, vol. 7, pp. 551–566, 2019. https://doi.org/10.1162/TACL_a_00282
Action Editor: Mamoru Komachi. Submission batch: 1/2019; Revision batch: 4/2019; Published 9/2019.
© 2019 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

The paper is organized into two main parts: data and metrics. In *data*, we first discuss how we create our data set, GMEG-Data (§3); then assemble the outputs of several GEC systems on this set (§4.1); and collect human judgments on the correctness of the outputs (§4.2). In *metrics*, we use these judgments to analyze how automatic evaluation metrics fare. We describe the existing metrics examined in this study (§5.1) and propose a new ensemble metric GMEG-Metric (§5.2). Finally, we evaluate and analyze metric performance on GMEG-Data and on the CoNLL-2014 Shared Task test set (§6).

2 Related Work

This paper represents an exploration of several components of GEC: corpora, metrics, and meta-evaluation. We summarize work in these areas.

2.1 Corpora

Prior work on evaluating GEC systems was performed on text written by primarily non-native English speakers, focusing on student essays by English language learners (ELLs). The NUS Corpus of Learner English (NUCLE) comprises essays written by mostly Chinese native speakers (Dahlmeier et al., 2013), and was the data set for the 2013 and 2014 CoNLL Shared Tasks in GEC (Ng et al., 2013, 2014). After the 2014 Shared Task, 16 additional references were released for that test set, 8 from each of Bryant and Ng (2015) and Sakaguchi et al. (2016). The Cambridge Learner Corpus First Certificate in English (FCE) data set includes essays for the B2 qualification exams (Yannakoudakis et al., 2011). The Johns Hopkins Fluency-Extended GUG corpus (JFLEG) contains text from the TOEFL exam (originally collected in the GUG corpus [Heilman et al., 2014]), with fluency corrections. *Fluency* corrections are rewrites needed to make a text sound natural to a native English speaker (Sakaguchi et al., 2016), in contrast to only making minimal corrections to grammatical errors as in FCE and NUCLE (Napoles et al., 2017). The Automatic Evaluation of Scientific Writing (AESW) shared task test set (Daudaravicius et al., 2016) has text from scientific publications written by proficient non-native and native English speakers, but is not widely used. We report the number of sentences and reference corrections for these four data sets in Table 1.

Name	Description	Size	# Refs.
CoNLL-14	Student writing	$1.3k$	$2 + 16$
FCE	Student writing	$2.7k$	1
JFLEG	Student writing	$1.5k$	4
AESW	Academic writing	$230k$	1
GMEG-Data	Formal/informal L1 and student writing	$8k$	4

Table 1: Summary of existing GEC test sets (size in number of sentences).

2.2 Metrics

The most commonly used automatic GEC metrics are MaxMatch (M^2) and GLEU. M^2 reports the F-score of edits over the optimal phrasal alignment between the candidate and the reference sentences (Dahlmeier and Ng, 2012). This was the official metric of the 2013 and 2014 Shared Tasks. The General Language Evaluation Understanding (GLEU) metric captures fluency rewrites in addition to grammatical corrections (Napoles et al., 2015, 2016a). It is an extension of BLEU (Papineni et al., 2002) that penalizes false negatives.

I-measure calculates the weighted accuracy of correction and detection, indicating how much better or worse a candidate system is than the original text (Felice and Briscoe, 2015). ERRANT is a rule-based error-type classifier (Bryant et al., 2017) that can be used as an evaluation metric by calculating the F-score of changes in a candidate text compared to a reference. It is the metric proposed for the 2019 shared task in GEC (Bryant et al., 2019). Other efforts have focused on reference-less metrics (Napoles et al., 2016b; Choshen and Abend, 2018b) and quality estimation (Chollampatt and Ng, 2018b).

2.3 Meta-evaluation of Metrics

Since 2006, the Workshop (now Conference) on Machine Translation (WMT) has conducted large-scale human evaluation of MT systems for its annual shared task (Koehn and Monz, 2006). The parallel Metrics track is a shared task for automatic MT metrics, evaluating performance against human judgments (Callison-Burch et al., 2008).

Researchers in GEC have adopted this practice following the CoNLL-2014 Shared Task on Grammatical Error Correction (henceforth *CoNLL-14*) (Ng et al., 2014), for which all results are publicly available, including the references and 13 system outputs. Grundkiewicz et al. (2015) and Napoles et al. (2015) simultaneously performed a human evaluation of the system outputs inspired by WMT. Heretofore, all work in GEC evaluation has been conducted on this data. Grunkdkiewicz et al. and Napoles et al. calculated the correlation of the human scores with M^2, I-measure, and BLEU, finding that M^2 moderately correlated with human judgments, I-measure had very weak negative correlation, and BLEU negatively correlated. Sakaguchi et al. (2016) analyzed the combination of available reference sets and metrics to identify the best evaluation configuration. Later, Chollampatt and Ng (2018c) re-examined the metrics, performing the first significance testing and calculating sentence-level correlation in addition to system-level correlations. They found no discernible difference between M^2 and GLEU and additionally determined that I-measure is actually a robust metric for sentence-level evaluation (also reported in Napoles et al. (2016b)).

Finally, Choshen and Abend (2018a) developed a methodology for automatically validating GEC metrics without human ratings by creating synthetic systems and calculating the correlation of a metric with the synthetic system ranking.

3 A Multi-domain Evaluation Set for GEC

The primary goal of this work is to enable GEC on broader types of writing beyond ELL student essays. First, we create a data set (GMEG-Data), and then use it to evaluate how robust existing metrics are (§5–6). Unlike prior test sets, GMEG-Data includes text written by native speakers[2] in formal and informal settings and ELLs from diverse backgrounds, and each sentence has been corrected by four professional annotators. We include 2,000 sentences from each of three sources:

- Informal Web posts (Yahoo Answers)
- Formal articles (Wikipedia)
- Student essays (FCE)

We selected these sources to represent diverse types of writing and to maintain contextual information when possible for future work in paragraph-level GEC. *Yahoo* contains paragraphs from Yahoo! Answers, written by users answering questions from other users,[3] and is very informal in terms of grammar, mechanics, and slang. *Wiki* has single sentences from Wikipedia, which is very formal and relatively well formed. We used the WikEd corpus (Grundkiewicz and Junczys-Dowmunt, 2014) as the source for Wikipedia sentences because it only contains text that has been changed by Wikipedia contributors and is therefore more likely to contain grammatical errors. We did not include additional context because a paragraph of Wikipedia text is less likely to contain multiple errors than the other two sources. Although the AESW is very large and encompasses formal text written by native and non-native English speakers, it overwhelmingly contains changes to punctuation (Flickinger et al., 2016) and has only been corrected once.

Finally, we chose paragraphs from *FCE* to represent ELL essays, because this continues to be an important domain in GEC research, and FCE represents a broad mix of 16 native languages. While the original FCE corpus has been corrected and there are already two other ELL test sets, we created new annotations for FCE because of shortcomings in those: The original FCE only has one reference; NUCLE represents a narrow sample of ELLs and topics; and JFLEG does not have any contextual information and was corrected by untrained, crowdsourced workers. This work does not use the original annotations released with the FCE and includes sentences from the training and test sets.[4]

3.1 Annotation Process

We randomly selected 2,000 sentences each from WikEd,[5] Yahoo, and the original FCE. For Yahoo and FCE, we chose whole paragraphs such that the total number of sentences was 2,000. All sentences have at least 5 tokens. Each sentence was corrected by four professional annotators, all native speakers of English, who were trained by two linguists. Annotators received detailed instructions with examples of common mistakes,

[2] We do not know the native language of all writers, but we assume the majority are native speakers of English based on qualitative evaluation.

[3] We include these categories: Arts/Humanities, Family/Relationships, Home, News, Politics, and Society/Culture.

[4] The source document IDs are included in the released data.

[5] 80% of the WikEd sentences contain named entities. Therefore we down-sampled these sentences so that only 60% of the sentences in our data set have named entities.

FCE	In ~~the present~~ **this day and age**, the ~~teachnology~~ **technology** is a ~~past~~ **fact** of life.
Wiki	**It's** ~~Unfortunate~~ **unfortunate** that you did not get to become an admin~~,~~ **;** I certainly did not ~~forsee~~ **foresee** such a mountain of opposition.
Yahoo	~~coz thats wat~~ **Because that's what** sustains us every now and then right from ~~the~~ birth.

Table 2: Examples of professionally corrected text from each domain in GMEG-Data. **Bolded** text indicates insertions and ~~strike-out~~ text was deleted. The edits correct grammatical, spelling, and fluency mistakes.

Name	Description	Sentences	Tokens/ sentence	Unchanged sentences	Edits/ sentence	Tokens/ edit	Type–token ratio
FCE	Student writing	1,936	18 ± 10	23%	2.8	1.3	0.11
Wikipedia	Formal	1,984	27 ± 13	19%	2.2	1.2	0.24
Yahoo	Informal	1,999	17 ± 11	50%	1.5	1.2	0.17
CoNLL-14	Student writing	1,312	23 ± 13	7%	2.2	1.4	0.10

Table 3: Description of test sets. *Edits/sentence* considers only changed sentences.

and they could ask questions of the trainers at any time. The same instructions were provided for each task and the complete instructions are provided with the annotated data. The trainers spot-checked their work for quality control.

Annotators were given entire paragraphs when available and instructed to correct all grammatical, spelling, and fluency mistakes. If two sentences had to be merged, annotators labeled those sentences as `MERGE`, and we concatenated pairs of sentences together if any annotator marked them `MERGE` (63 pairs). When annotators could not correct a sentence, they assigned the label `FRAGMENT` if that sentence represented a fragment of text or `NOT ANNOTATABLE` for other issues. Fifteen sentences were removed because all annotators marked them as `NOT ANNOTATABLE` or `FRAGMENT`. Table 2 shows example edits from each domain.

3.2 Annotation Analysis

We first calculated descriptive statistics for each of the three domains and CoNLL-14[6] as a point of comparison. We ran ERRANT over the references and original sentences to obtain alignments and error type categories. Table 3 summarizes the extent of changes in each domain.

In all domains, the average length of each edit is fairly consistent, which can be attributed to using the edit spans identified by ERRANT. The portion of sentences needing corrections varies across domains. The non-native corpora of student writing, FCE and CoNLL-14, have fewer than half as many correct sentences as Yahoo, which was written by native speakers. They also have the least amount of lexical variety, captured by the type–token ratio. Wikipedia, on the other hand, has fewer unchanged sentences than FCE, presumably because the sentences were drawn from WikEd, which was intended to be a corpus of parallel edits. It also has the most lexical variety, which may be due to the high frequency of named entities. Yahoo, the most informal domain, contains slang terms and non-conventional mechanics. Somewhat surprisingly, Yahoo has the most uncorrected sentences but also the fewest number of edits per corrected sentence. Even though the annotators received the same guidelines for all domains and were not instructed to preserve the register of the text, they appear to have preserved the more informal style by not making extensive rewrites to the text.

Figure 1 shows the distribution of the most frequent error types classified by ERRANT, not including the `OTHER` category.[7] The majority of errors in Yahoo and Wiki are punctuation, orthography, and spelling, which is not surprising as these sentences were written by native speakers. Table 2 shows examples of the types of spelling and punctuation mistakes typical in these corpora. Yahoo has non-standard spelling, capitalization, and punctuation, typically found in less formal

[6] All experiments in this paper use the original two CoNLL-14 references.

[7] `OTHER` is assigned to any error not captured by the ERRANT rules. It is prevalent: The most frequent category in CoNLL-14, the second most in FCE and Wiki, and the third most on Yahoo.

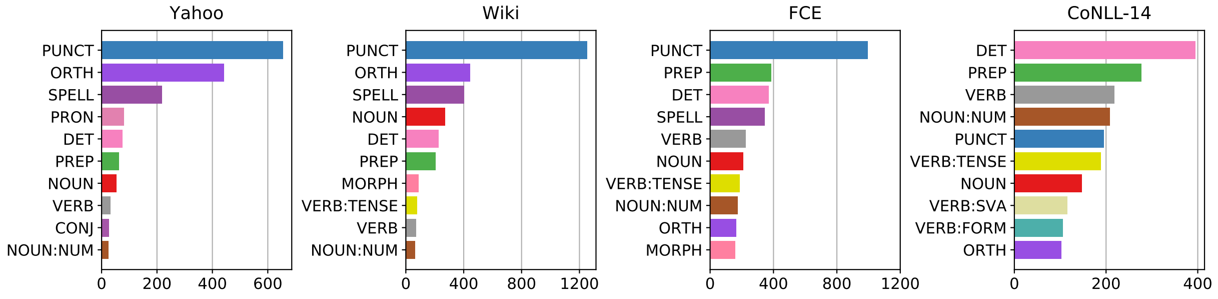


Figure 1: Distribution of the ten most frequent edit types by domain.

text, and it is interesting to note that the relative frequency of each error type in these domains is similar in spite of these differences. FCE contains more grammatical errors, including prepositions, determiners, verbs, and nouns.

The distribution of errors in GMEG-Data vastly differs from CoNLL-14. In CoNLL-14, the error types are more evenly distributed, whereas PUNCT (punctuation errors) are the most frequent mistakes in our annotations. The difference can be attributed to how the annotations were collected: the annotators of CoNLL-14 had to categorize their corrections, focusing the annotations to fit into each category, and thereby making the errors easier to categorize. For our corrections, annotators had to select a span and type a replacement but did not label the error type, which could have affected the performance of ERRANT. Although both FCE and CoNLL-14 contain essays written by ELLs, the texts are quite different: The CoNLL-14 essays were written by a fairly homogeneous group of students at the same university, with similar English-language proficiency levels, and with fewer native languages than FCE. FCE represents writing from students with more than a dozen different native languages from different geographic regions. While punctuation is the most frequent error in each of our new domains, the majority of corrections are to other types of errors, in contrast to the AESW corpus, which has punctuation accounting for more than half of the edits (Flickinger et al., 2016).

The source data for FCE and Wiki both have corrections, and we compare this new set of annotations to the original ones. GMEG-Data has more corrections per sentence than the original corpora: an average of 2.2 corrections/sentence for FCE compared with 1.7, and in Wiki, 1.9 compared with 1.5. We additionally calculate the M^2 precision and recall of the original annotations against the new corrections to understand the extent of overlap between the annotation sets (Table 4). For FCE, $P = 64.8$, which is comparable to the precision of the GMEG-Data references against each other, but recall is much lower. This suggests that many of the original FCE corrections are included in GMEG-Data. In contrast, the original Wiki annotations have much lower precision and recall (40.9 and 33.2, respectively). WikEd was taken from the Wikipedia revision history and we surmise that the original corrections contain many changes not related to GEC mistakes such as *Copenhagen* → *Østerbro*. Considering just GMEG-Data, the high mean precision and recall indicates a high level of consistency between the different references.

	Original		GMEG-Data	
Domain	P	R	P	R
FCE	64.8	49.6	67.3 ± 3.4	63.2 ± 4.7
Wiki	40.9	33.2	61.6 ± 1.1	59.4 ± 4.3

Table 4: The M^2 precision and recall of the original set of corrections against the GMEG-Data references and the mean scores of each GMEG-Data reference against the other three references.

4 Human Evaluation of GEC Systems

Now that we have a new test set in place with reference annotations, the next step is to correct the source sentences with GEC systems. We will then perform a human evaluation of the system outputs. This provides the setup for evaluating automatic metrics (conducted in §5). Correlation with human rankings has been used to evaluate GEC metrics on the CoNLL-2014 Shared Task results (Grundkiewicz et al., 2015; Napoles et al., 2015; Sakaguchi et al., 2016; Chollampatt and Ng, 2018c). Following evaluation methodologies established in WMT, previous work

Name	Architecture	Tokens	Data size
AMU	SMT	word	Large
LSTM	LSTM	BPE	Large
LSTM-R	LSTM	BPE	Large
Marian	deepGRU	BPE	Small
NUS	CNN	BPE	Medium
Transf.	Transformer	BPE	Large

Table 5: Different types of system architectures. *Data size* indicates the number of parallel training sentences. Small: <1M, Medium: 1M–2M, Large: >2M.

collected human ratings of 13 system outputs from CoNLL-14 and calculated the correlation between automatic metric scores and those human judgments. This section describes assembling ground-truth human judgments of GMEG-Data. We first outline the GEC systems included and then describe how we conduct human evaluation of the system outputs.

4.1 GEC Systems

We select 6 GEC systems that differ in 3 aspects: the type of system (statistical or neural), the neural network architecture, and the amount of data used for training. We give only a brief overview of the systems below (summarized in Table 5) because the goal of this paper is not to identify the best GEC system, and furthermore the systems are trained on a combination of public data sets and propriety data.

AMU A statistical MT model trained using a modified version of the Moses toolkit (Koehn et al., 2007). We use the pre-trained model published in Junczys-Dowmunt and Grundkiewicz (2016).

LSTM A RNN-based sequence-to-sequence neural network with bi-directional encoder and LSTM units, trained with the OpenNMT-py toolkit (Klein et al., 2018). We train a second high-recall system using the same architecture, but changing the data sampling strategy (called **LSTM-R**).

Marian A RNN-based sequence-to-sequence neural network with deep-transition architecture (Barone et al., 2017) trained with the Marian toolkit (Junczys-Dowmunt et al., 2018). We use the WMT17 system parameters (Sennrich et al., 2017), excluding ensembles and left-to-right re-ranking.

NUS A multi-layer convolutional sequence-to-sequence neural network trained with the Fairseq-py toolkit (Gehring et al., 2017). We use the pre-trained model of Chollampatt and Ng (2018a).

Transformer A transformer (Vaswani et al., 2017) neural network trained using the Fairseq-py toolkit and the parameters proposed for the IWSLT '14 MT task.[8]

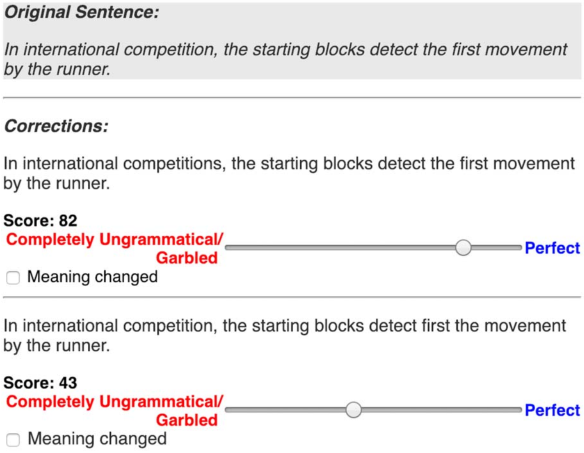
4.2 Human Judgments of GEC

We evaluate all 6 GEC systems described above as well as 3 additional ''systems'' to define a baseline (**Source**: the unaltered input sentence), lower-bound (**Source+error**: the source sentence with 1–2 errors inserted[9]), and upper-bound (**Reference**: a randomly chosen human correction). Our evaluation includes GMEG-Data as well as CoNLL-14, for comparison with earlier work. Unlike previous human evaluations (Grundkiewicz et al., 2015; Napoles et al., 2015), we collect judgments on the entire set of sentences from both corpora. Both human and automatic evaluation are performed over the complete set of sentences, thus addressing Choshen and Abend's (2018a) critique of prior work using inconsistent samples for human and automatic evaluation.

Several methodologies have been proposed for human evaluation of system outputs. For this task, we use a hybrid approach that combines judgments on a continuous scale with relative ranking (*partial ranking with scalars* or PRWS). PRWS was advocated for in EASL (Sakaguchi and Van Durme, 2018) and RankME (Novikova et al., 2018), two recent works that investigated reliable methods for collecting human ratings of competing systems' outputs. Those studies both found PRWS to be more reliable than the direct assessment framework used in WMT (Bojar et al., 2016), the relative-ranking methodology formerly applied by WMT (Callison-Burch et al., 2007), and earlier GEC human evaluation work (Grundkiewicz et al., 2015; Napoles et al., 2015). Unlike relative ranking, PRWS does not explicitly ask raters to rank the sentences, although a ranking can be inferred from the relative scores. Raters

[8] `https://github.com/pytorch/fairseq/tree/master/examples/translation`

[9] These sentences and an explanation of the rule-based error-insertion method are included with the released data set.



Original Sentence:

In international competition, the starting blocks detect the first movement by the runner.

Corrections:

In international competitions, the starting blocks detect the first movement by the runner.

Score: 82

Completely Ungrammatical/ Garbled
Perfect

□ Meaning changed

In international competition, the starting blocks detect first the movement by the runner.

Score: 43

Completely Ungrammatical/ Garbled
Perfect

□ Meaning changed

Figure 2: Interface for collecting human ratings.

implicitly adjust their scores for each system to be relative to the other systems and thereby the numeric scores are more discriminative.

In the PRWS framework, human participants read the original sentence and up to 5 corrected versions of that sentence. For each correction, they drag the handle of a slider bar to their judgment on a scale from *Completely ungrammatical/Garbled* to *Perfect*. The position on the slider bar corresponds to a score between 0 and 100. We also include a checkbox to indicate whether a correction altered the meaning of the sentence. Our PRWS interface is shown in Figure 2.

We performed annotation on Amazon Mechanical Turk, and showed workers one group of judgments on the screen at a time, with 20 screens per HIT following the direct-assessment framework (Graham et al., 2015), allowing for more robust quality control. Every sentence generated by every system was scored, with duplicate system outputs collapsed. Domains were separated so that no worker judged sentences from different domains on the same day. Each item was evaluated by 8 participants, limited to be in the US with an acceptance rate of 98% and $\geq$ 500 HITs completed, and paid \$2 per HIT. If a worker did not assign a lower score to *Source+error* than the *Source* sentence at least 70% of the time, we excluded their work. In total, 34 out of 771 workers were excluded from the task.

4.3 Analysis of Human Judgments

From the judgments, we calculate human scores for each sentence and for each system. The score of each candidate sentence is the mean of the 8 human ratings of that sentence, and the system score is the mean of all candidate sentence scores produced by that system. We use the mean score instead of the relative rank for this study because an automatic metric must be able to evaluate a single system by providing a numeric score (0–100 in our case). Table 6 displays the scores of each system by domain. Across domains, the difference between the scores of the source and reference is around 7 points, except for FCE, which has more than twice as wide a difference (15 points). The highest performing system by domain makes up about half of the performance gap between the source and reference, however, the best systems change far fewer sentences than the reference (13% points fewer on average). The overall ranking of systems changes based on domain, with all systems outperforming the *Source* on FCE and CoNLL-14, but *Source* is judged higher than at least 2 systems in the other domains. Transformer has the lowest score in Wiki and Yahoo, likely because it is more aggressive than the other systems: It leaves the fewest number sentences of unchanged and changes the meaning of 3.7% of sentences, more than every system but Source+error. NUS has the next most sentences with meaning changes (1.1%) and all other systems change meaning in fewer than 1% of sentences.

We calculate the reliability of the human ratings by splitting the 8 judgments into 2 groups of 4 and reporting the mean correlation between each group (Table 7). At the system level, correlation is very strong (> 0.9), demonstrating the consistency of the annotations collected with PRWS and further suggesting that even 4 judgments yield robust scores at the system level. At the sentence level, correlation is more variable across domains. CoNLL-14 and Wiki have the lowest correlations (0.3–0.4), and FCE and Yahoo have moderate correlation (0.5–0.6). More careful scrutiny is needed to understand the differences between domains at the sentence level.

5 Evaluating GEC Metrics

To evaluate existing metrics, we calculate the correlation between the human scores (collected in §4.2) and the scores assigned by an automatic metric. We will examine existing metrics (§5.1) and a new ensemble metric that we propose in §5.2.

	Mean human score				Percent of sentences unchanged			
System	FCE	Wiki	Yahoo	CoNLL	FCE	Wiki	Yahoo	CoNLL
Ref.	83.4	82.1	82.3	79.9	23.1	18.9	49.7	7.3
AMU	*70.7*	76.0	74.5	74.0*	**60.3**	**70.3**	**75.9**	47.6*
LSTM	74.3	77.7	**78.5**	**75.9**	34.3	40.7	49.9	28.1
LSTM-R	74.4	**78.2**	78.2	75.6	30.7	34.4	42.4	28.2
Marian	**77.0***	75.5	76.6	*71.2*	40.0*	65.6	64.5	53.9
NUS	74.0	75.7	75.9	73.4*	52.3	67.5	65.0	**55.6***
Transf.	73.9*	*71.5*	*72.2*	75.6*	*28.6**	*15.0*	*25.9*	*19.8**
Source	68.1	75.9	75.1	70.2	100.0	100.0	100.0	100.0
Source+error	46.0	62.4	47.5	53.8	0.1	0.4	0.1	0.1

Table 6: The mean score assigned to each system and the percent of sentences left unchanged (*No change*). The highest value is in **bold**, the lowest in *italics*. * indicates that the model was trained on in-domain data for that test set.

	System		Sentence	
Domain	r	ρ	r	ρ
FCE	0.992	0.914	0.603	0.588
Wiki	0.994	0.907	0.376	0.285
Yahoo	0.988	0.990	0.599	0.543
CoNLL-14	0.981	0.914	0.399	0.387

Table 7: Pearson (r) and Spearman (ρ) correlation between the scores of 2 groups of 4 human judges at the sentence- and system-level, averaged over 100 samples.

5.1 Existing Metrics

We include the following GEC and MT metrics in our analysis:

GLEU With the default order n-grams, $n = 1.4$.

charGLEU GLEU modified to use character n-grams ($n = 1.5$).

I-measure With the `-nomix` option to speed up computation time.

MaxMatch (M^2) We report $\mathrm{M}^2_{0.5}$ (with the default parameter $\beta = 0.5$) and $\mathrm{M}^2_{0.2}$ ($\beta = 0.2$), which was shown to have stronger correlation with human ratings (Grundkiewicz et al., 2015).

ERRANT The average $F_{0.5}$ or $F_{0.2}$ score of the 24 error categories assigned by ERRANT, without the finer grained distinction between Missing, Replacement, or Unnecessary edits.

CHRF++ A top metric at the WMT Metrics Shared Tasks (Ma et al., 2018; Bojar et al., 2017) that reports the F_2 score of character n-grams ($n = 1..6$) and token n-grams ($n = 1, 2$) (Popović, 2017).

5.2 Ensemble Metric

We propose a new scorer, GMEG-Metric, that is an ensemble of the existing metrics. The motivation is twofold. First, different metrics capture different aspects of the correction. For instance, ERRANT precisely represents the precision and recall by specific error type and, arguably, error types should not be equally weighted (correcting an agreement error is likely more impactful than correcting a punctuation error). GLEU, on the other hand, captures the fluency of corrections by comparing higher-order n-grams. Second, different domains of text have different types of errors and corrections, and therefore we suspect that the reliability of a metric may not be constant across different domains. For example, the CoNLL-14 data set, on which previous metrics have been evaluated, has very few punctuation and spelling corrections, unlike the three domains included in GMEG-Data (as shown in Figure 1). Additionally, by training a supervised metric using data from a range of domains and high-performing GEC system architectures, we provide a way to evaluate black-box systems for which the training regime and data are unknown.

To facilitate system development, ideally a metric should incorporate information about performance on specific error types, fluency, recall, and precision. In the field of MT, metrics have been proposed that use a regression model, trained on data from multiple language pairs and domains, to predict the sentence-level human scores. Some of these models use basic statistics such as precision and recall of character and word n-grams (Stanojevic and Sima'an, 2014), while

	FCE		Wiki		Yahoo	
Metric	r	ρ	r	ρ	r	ρ
CHRF++	0.639	0.733	**0.972**	**0.943**	0.791	0.687
ERRANT$_{0.2}$	0.828	0.779	0.453	0.596	0.596	0.643
ERRANT$_{0.5}$	0.919	0.887	0.401	0.555	0.532	0.601
GLEU	0.838	0.813	0.426	0.538	0.740	0.775
charGLEU	**0.959**	**0.932**	0.644	0.732	0.846	0.835
I-measure	0.819	0.839	0.854	0.875	**0.915**	**0.900**
M$^2_{0.2}$	0.852	0.846	0.548	0.680	0.717	0.836
M$^2_{0.5}$	0.860	0.849	0.346	0.552	0.580	0.699
GMEG-Metric	**0.984**	**0.950**	**0.982**	**0.967**	**0.940**	**0.931**
Human	0.992	0.931	0.994	0.907	0.988	0.990

Table 8: Correlation of metrics with human scores on the test set (Pearson's r and Spearman's ρ). *Human* is the correlation between two human scores, each containing 4 randomly selected non-overlapping judgments, averaged over 100 iterations. For the automatic metrics, correlation is calculated against all 8 judgments.

others use the scores of other metrics as input features (Yu et al., 2015; Ma et al., 2017).

The ensemble scorer we propose for GEC is a ridge regression model ($\alpha = 0.001$, determined with cross-validation)[10] trained to predict the system-level human scores. The model uses as features the precision and recall values calculated by ERRANT, CHRF, and M^2; the statistics calculated by I-measure, not including F-scores; and the overall GLEU and charGLEU scores (73 features in total). We train a single model on the combined data from all domains, with each system as an instance.

To counter the small number of data points available for training, we generate 674 new synthetic systems for each domain from the existing 6 GEC systems. Synthetic systems are created following the hybrid super-sampling approach of Graham and Liu (2016), which was used to generate new systems for evaluating MT metrics on language pairs with a small number of systems. We sample sentences from pairs of systems (not including Reference, Source, and Source+error), altering the percentage of sentences from each system from 10% to 90%.[11] The system-level score for each synthetic system is the mean of the human scores of each sentence it contains.

We split the annotated data in half, so that there are approximately 1,000 sentences from each domain for training/development and 1,000 for testing. GMEG-Metric is trained on the development set with cross-validation.

6 System-level Correlation with Human Judgments

Each metric is evaluated on $3k$ sentences in the test split. We report the Pearson correlation and the Spearman rank coefficients between the system scores predicted by each metric and the ground-truth scores from the human rating. The Spearman coefficient is appropriate for differentiating between two systems, although it too harshly penalizes metrics that change the order of systems with similar performance (Macháček and Bojar, 2013), so we also report the Pearson correlation. Because we only have a small number of systems, and the correlation can change based on the inclusion or exclusion of one additional data point, we include the artificial systems (§5.2) for more robust correlation calculations. The upper- and lower-bounds (*Reference* and *Source+error*) could artificially inflate the correlation values and are therefore not included in any of the correlation calculations. The results on the test set for all metrics and domains are in Table 8. Following Graham and Baldwin (2014), we apply the Williams statistical test[12] over the Pearson correlation values to find the best metric. Figure 3 shows the matrix of p-values for each pair of metrics.

[10] Using the scikit-learn toolkit (Pedregosa et al., 2011).

[11] The method for producing synthetic systems is included with the released data and code.

[12] https://github.com/ygraham/significance-williams

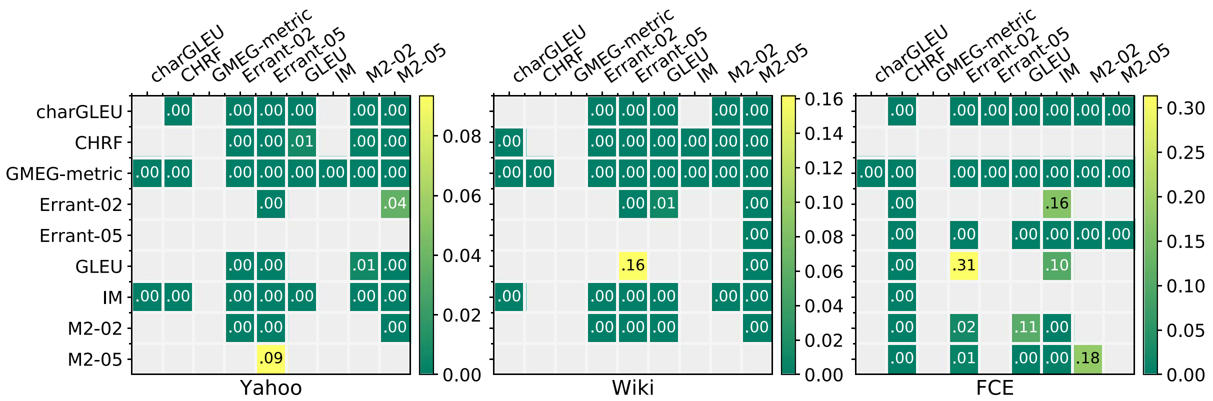


Figure 3: Results of Williams significance test. The p-value of the cell (i, j) indicates that metric i is significantly better than metric j. The best metric on each domain is identified by a completely gray column, meaning no other metric was significantly ($p < 0.05$) better.

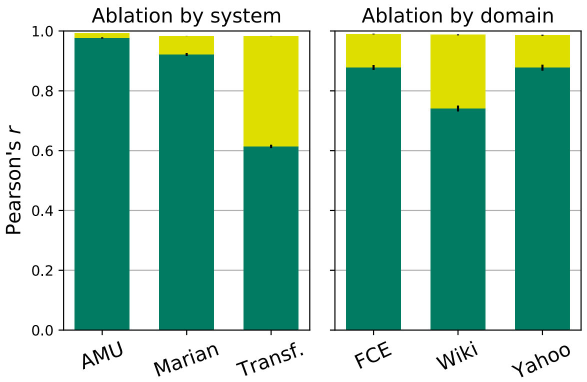


Figure 4: Results of ablation by system and by domain. The yellow bar indicates performance when the domain is seen in training, and the green bar the ablation result.

The best metric in all domains is GMEG-Metric. In some instances, GMEG-Metric has stronger correlation than Human, which we explain because GMEG-Metric was trained to predict the mean of 8 human ratings, whereas Human is the correlation between 2 groups of 4 human ratings. The previously proposed metrics have inconsistent performance across domains. For instance, I-measure is the second best metric in the Yahoo domain (only GMEG-Metric is better) but the second worst metric in FCE (only CHRF is worse). Not including GMEG-Metric, I-measure is statistically best on Yahoo, CHRF is the best on Wiki, and charGLEU is the best on FCE.

Character-level GLEU (charGLEU) is statistically better than word-level GLEU, which can be explained by the high number of spelling and orthographic corrections present in the three data sets. Furthermore, charGLEU is better than the M^2 metrics on all three domains. On the domains with fewer errors per sentence, Wiki and Yahoo, precision becomes more important as signaled by the statistically higher correlation of $M^2_{0.2}$ compared with $M^2_{0.5}$. ERRANT$_{0.5}$ and $M^2_{0.5}$ have high correlations only on FCE, which has a higher error rate and a more even distribution across error types. In contrast, CHRF++, which operates on n-grams, has the lowest correlation on FCE. It inflates the scores of systems that have low edit recall but high n-gram recall because there is naturally a high n-gram overlap between the candidate and reference text.

When training on all domains, the ensemble scorer performs the best across domains according to the significance test, capturing the relative importance of different metrics. Given the variability in error distributions and metric performance across the three domains, we recommend using the trained ensemble scorer for future system development.

6.1 Ablation Studies

To understand how each of the domains and systems impacts the overall performance of the ensemble scorer, we perform a series of ablation studies on the development set, reporting the mean performance over 50 folds of cross-validation with a 75/25 train/test split. We report the Pearson correlations[13] in Figure 4.

By system We ablate each system and the corresponding group of synthetic systems that were sampled from that system. We calculate correlations on a 25% split of each subset of the data, and report results for the three systems

[13] We obtained very similar results with the Spearman correlation.

Domain	All	Ablation
FCE	0.962	0.915
Wiki	0.973	0.507
Yahoo	0.969	0.468
CoNLL-14	0.978	−0.632

Table 9: The Pearson correlation of GMEG-Metric trained on all four domains, averaged over 50 cross-validation folds. *All* reports the score for the model trained on all domains and *Ablation* reflects the correlation when that domain is ablated.

that showed a significant change (Figure 4). The correlations remain very high for all systems, except when ablating Transformer. This system was judged the lowest by humans on Wiki and Yahoo and proposed the most meaning-changing corrections (3.7% of sentences). Such corrections include deletions or substitutions of named entities (e.g., *''He was born Mads ~~Dittmann Mikkelsen~~* ***in*** *in 1965''*), and of infrequent words (e.g., *''Truvannamalai is a ~~cosmopolitan~~* ***city*** *city''* and *''Many ~~non-retail~~* ***foreign*** *offices are closed''*). A limitation of the existing metrics is they do not penalize such changes as much as humans do. Meaning-changing corrections are less prevalent in the output of the other systems, and so when Transformer is ablated, these types of changes are no longer present in the training data and the model no longer penalizes them. We include Transformer for training GMEG-Metric so it is robust to unseen systems that make aggressive, meaning-altering changes, in addition to competitive systems that do not exhibit this behavior. For future work, we propose augmenting the GMEG-Metric with a feature for detecting meaning changing edits.

By domain Next, we ablate systems from each domain and report the results of testing on a 25% split of the domain data over 50 cross-validation folds. As expected, performance drops when a domain is not seen in training, although the correlation is still high. Even the lowest result, 0.74 Pearson correlation on Wiki, indicates a strong correlation that is higher than the correlation reported in Table 8 for all but two of the existing metrics, suggesting the metric is robust across the three domains.

6.2 Contrastive Analysis on CoNLL-14

Finally, to draw a comparison with previous GEC metrics research, we run the same experiments with the test set from the CoNLL-2014 Shared Task, reporting cross-validation results over the GMEG-Data development and the CoNLL-2014 test set.[14] The two references provided in the shared task are used as the gold standard with the automatic metrics. The systems in this work are significantly better than the systems that participated in the 2014 Shared Task: The best shared-task system had $M^2_{0.5} = 37.3$ whereas the $M^2_{0.5}$ of our systems ranges from 42.1 to 51.6. Figure 5 shows the correlations, which differ substantially from the correlations we found on the new domains (Table 8). Of note, GMEG-Metric, CHRF, and I-measure negatively correlate with the human score, while all metrics have positive correlation on the new domains. However, the correlations for GLEU, I-measure, and $M^2_{0.5}$ are similar to those that have been reported in previous work (Grundkiewicz et al., 2015; Napoles et al., 2015), supporting the findings of those studies and further suggesting that these metrics score improved systems as reliably as the state-of-the-art from 2014. This also supports our annotation framework, which yields results consistent with results reported in previous studies (in addition to being internally consistent; see §4.3).

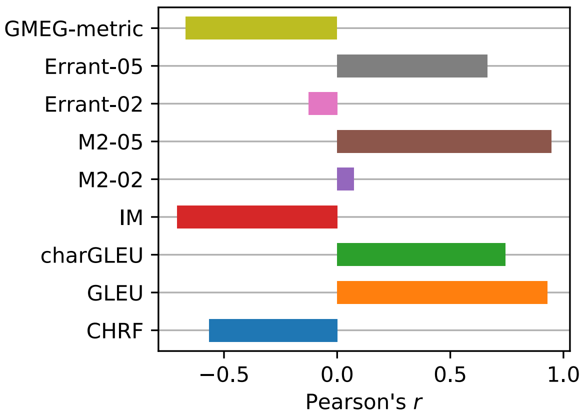


Figure 5: Correlation of automatic metrics on new systems (including artificial) on the CoNLL-14 test set.

GMEG-Metric, which was trained on the combined data of the other three domains, has a moderate negative correlation on the CoNLL-14 test set. This result is yet another signal that this data set is significantly different from the new domains, since the ablation study shows GMEG-Metric has stable performance on Yahoo, Wiki, and FCE. Notably, CoNLL-14 has fewer

[14] We use cross-validation because the CoNLL-2014 test set is smaller, with only 1,312 sentences.

punctuation and spelling corrections, as well as a different error-type distribution compared to FCE, the other data set of non-native text (Figure 1). Of the metrics with strong positive correlation on CoNLL-14, charGLEU is the only one that is stable and performs relatively well on the other three domains.

To support researchers who want to optimize their GEC systems on the CoNLL-14 test set, we train the ensemble including this narrow domain and report cross-validation results in Table 9. Although the GMEG-Metric has very strong correlation with human scores when trained on all four domains, performance drops significantly when Wiki, Yahoo, or CoNLL-14 are ablated. FCE is the only domain that still has strong performance when unseen in training. Without explicit domain adaptation, the supervised metric is biased towards the domains present in the training data. These results demonstrate the expressiveness of GMEG-Metric to new domains of text, including domains that have different error distributions. In order to evaluate GMEG-Metric on a new domain, a new model should be retrained with representative data.

7 Conclusions and Future Work

In this paper, we take the first major step forward towards robust grammatical error correction across multiple domains. Prior work in GEC has focused almost exclusively on correcting errors by ELL writers, which has led to systems optimized for that demographic and metrics optimized to the few data sets available. However, ELL writers are just one of the many potential beneficiaries of GEC feedback. We advance the field by releasing a new multi-domain, multiple-reference data set (GMEG-Data) encompassing two new domains of text written by native speakers. Each domain is large enough to be a stand-alone evaluation set, having more sentences than prior multiple-reference GEC test sets. The data set additionally includes corrections by 6 current GEC systems and 8 human ratings per correction. We also release a pre-trained ensemble scorer (GMEG-Metric) that we have shown to be the best metric on all three domains. GMEG-Metric is flexible and robust when retrained on new domains of text: When including CoNLL-14 in training, the metric has very strong correlation with human judgments, compared with a strong negative correlation when it is not included in the training data. To adapt the metric to a new domain, human judgments are necessary for no more than 1,000 sentences from 6 ‘‘real’’ systems.

With this work we are able to draw three important conclusions:

1. Metrics used for ELL domains (i.e., CoNLL-14) are not the most reliable on other domains. In fact, the best singular metrics are I-Measure and a character-version of GLEU developed in this work.

2. Our ensemble metric (GMEG-Metric) has the highest correlation with human judgments and we recommend GEC practitioners use this framework as the field expands to new domains.

3. The introduction of a new GEC evaluation set should be accompanied by a meta-evaluation of the automatic metrics on that data.

There are several avenues for future work. By sharing all of the collected data, system outputs, and annotations,[15] we lay the groundwork for further metric development. This data set also allows the field to experiment with domain adaptation methods for the first time: For instance, tailoring neural GEC models trained on ELL data to one of our native domains, or tuning GMEG-Metric on a small amount of data to work robustly across a multitude of domains. Other research avenues include paragraph-level GEC and sentence-level quality evaluation.

Acknowledgments

The authors would like to thank Yahoo Research for making their data available. We are grateful to our colleagues for their assistance in this project: Dimitrios Alikaniotis, Maksym Bevza, Phu Mon Htut, Claudia Leacock, Dmitry Lider, Kostiantyn Omelianchuk, and Igor Tytyk. We thank Mamoru Komachi, Lillian Lee, and the three anonymous reviewers for their helpful feedback.

References

Miceli Barone, Antonio Valerio, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2017. Deep architectures for neural

[15] https://github.com/grammarly/GMEG

machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 99–107, Copenhagen.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 Metrics Shared Task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 489–513, Copenhagen.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 Metrics Shared Task. In *Proceedings of the First Conference on Machine Translation*, pages 199–231, Berlin.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications*. Florence.

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)*, pages 793–805, Vancouver.

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.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, OH.

Shamil Chollampatt and Hwee Tou Ng. 2018a. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5755–5762.

Shamil Chollampatt and Hwee Tou Ng. 2018b. Neural quality estimation of grammatical error correction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2528–2539, Brussels.

Shamil Chollampatt and Hwee Tou Ng. 2018c. A reassessment of reference-based grammatical error correction metrics. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2730–2741, Santa Fe, NM.

Leshem Choshen and Omri Abend. 2018a. Automatic metric validation for grammatical error correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1372–1382, Melbourne.

Leshem Choshen and Omri Abend. 2018b. Reference-less measure of faithfulness for grammatical error correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 124–129, New Orleans, LA.

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.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, GA.

Vidas Daudaravicius, Rafael E. Banchs, Elena Volodina, and Courtney Napoles. 2016. A report on the Automatic Evaluation of Scientific

Writing shared task. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 53–62, San Diego, CA.

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

Dan Flickinger, Michael Goodman, and Woodley Packard. 2016. UW-Stanford system description for AESW 2016 shared task on grammatical error detection. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 105–111, San Diego, CA.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, Sydney.

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 (EMNLP)*, pages 172–176, Doha.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, Denver, CO.

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, CA.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2014. The WikEd error corpus: A corpus of corrective Wikipedia edits and its application to grammatical error correction. In *International Conference on Natural Language Processing*, pages 478–490.

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

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

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1546–1556, Austin, TX.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 177–184.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association* *for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York, NY.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 Metrics Shared Task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 682–701, Brussels.

Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. 2017. Blend: A novel combined MT metric based on direct assessment–CASICT-DCU submission to WMT17 Metrics Task. In *Proceedings of the Second Conference on Machine Translation*, pages 598–603, Copenhagen.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 Metrics Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel R. Tetreault. 2016a. GLEU without tuning. *CoRR*, cs.CL/1605.02592v1.

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*, pages 2109–2115, Austin, TX.

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

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, MD.

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.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78.

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, PA.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Maja Popović. 2017. CHRF++: Words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen.

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 and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 389–399, Copenhagen.

Milos Stanojevic and Khalil Sima'an. 2014. BEER: BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, MD.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017, Attention is all you need, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new data set and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189.

Hui Yu, Qingsong Ma, Xiaofeng Wu, and Qun Liu. 2015. CASICT-DCU participation in WMT2015 metrics task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 417–421, Lisbon.