How Good (really) are Grammatical Error Correction Systems?

Alla Rozovskaya
Queens College, CUNY
arozovskaya@qc.cuny.edu

Dan Roth
University of Pennsylvania
danroth@seas.upenn.edu

Abstract

Standard evaluations of Grammatical Error Correction (GEC) systems make use of a fixed reference text generated relative to the original text; they show, even when using multiple references, that we have a long way to go. This analysis paper studies the performance of GEC systems relative to *closest-gold* – a gold reference text created relative to the *output* of a system. Surprisingly, we show that the real performance is 20-40 points better than standard evaluations show. Moreover, the performance remains high even when considering any of the top-10 hypotheses produced by a system. Importantly, the type of mistakes corrected by lower-ranked hypotheses differs in interesting ways from the top one, providing an opportunity to focus on a range of errors – local spelling and grammar edits vs. more complex lexical improvements. Our study shows these results in English and Russian, and thus provides a preliminary proposal for a more realistic evaluation of GEC systems.

1 Introduction

Grammatical Error Correction (GEC) systems are typically evaluated using reference-based evaluation measures. This is common in language generation tasks, where the system output is compared against a set of gold references, such as the set of correct translations in Machine Translation or the set of valid corrections for a source sentence in GEC. Importantly, the references are generated relative to the original text and are independent of the system outputs. In GEC, the space of valid outputs for a given source sentence is very large, making it extremely difficult to evaluate. Specifically, *reference-based evaluations (most GEC datasets contain one reference correction) are known to underestimate system performance* (Chodorow et al., 2012; Felice and Briscoe, 2015; Bryant and Ng, 2015). Bryant and Ng (2015) showed that using two references is better than one, and the results improve further with more references; however, they used references that were generated relative to the original text. Choshen and Abend (2018b) further demonstrated that the issue can be only slightly alleviated but not completely solved by increasing the number of references. This is because many errors have a long-tailed distribution on valid corrections. One can expect that as GEC systems mature and manage to address more complex errors, the underestimation of their performance will be further exacerbated.

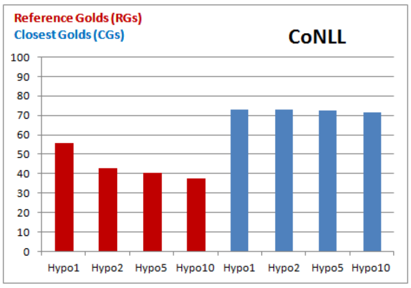

Figure 1: Performance by hypothesis rank (F-score) against Reference Gold (RG) vs. Closest Golds (CGs) generated for each hypothesis. Observe dramatic drop in performance between top hypothesis and the rest in RG evaluation, vs. stability in CG evaluation; and, large gaps between scores even for the top hypotheses in RG vs. CG evaluations.

Choshen and Abend (2018b) discuss another consequence of having a large space of valid corrections, pertaining to training. They show that *GEC systems have a strong tendency to under-correct*, due to using the one-reference-gold approach for tuning (and training) the systems. Essentially, due to the low likelihood of a system's proposed change being matched to gold, GEC systems are discouraged from proposing corrections, and propose far fewer corrections compared to humans. The under-correction phenomenon is more pronounced for errors with a large number

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2686–2698
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

of correction candidates. For example, mistakes on closed-class words, e.g. errors in determiners, where the number of valid corrections is small, suffer from under-correction to a lesser extent than mistakes in word choice. As a result, current systems generally prefer to make small targeted changes on closed-class errors.

We further study the effects of having a large space of valid corrections on GEC system development and automatic evaluation. Given a potentially erroneous sentence, we assume there is a space of gold references corresponding to it. Evaluation is typically done by drawing one gold from this set. We refer to this as *reference gold* (RG). We generate a new gold that is as close as possible to the system output (*hypothesis*), by correcting the hypothesis itself, instead of the original text. We call it *closest gold* (CG). We show by how much performance is really underestimated when a reference gold is used instead of the closest one, and claim that the latter should reflect the true performance of the system. We use a ranked 10-best list of hypotheses for a given source sentence, produced by state-of-the-art GEC systems on two English and two Russian GEC datasets. We generate CGs for hypotheses at different ranks.

Our findings are as follows. First, evaluation against RGs shows a large performance gap between the top hypothesis and the rest. We show that the reason for this is that lower-ranked hypotheses propose more diverse changes, including lexical changes, that have a lower chance to match RGs. In contrast, evaluation against CGs reveals that qualitatively, there is very little degradation in the hypotheses, when considering the top-10 list. While RG evaluation reveals severe drops in F-score and, in particular, precision, we find that relative to CGs quality does not substantially degrade. This is illustrated in Figure 1 for one of the datasets; we show more results in Section 4.1.

Second, contrary to the observation made by Choshen and Abend (2018a) about GEC systems being disinsentivized to propose corrections, we find that it only applies to the top-ranked hypothesis.[1] Moreover, the number of proposed edits increases steadily with the hypothesis rank.

We further evaluate the output by computing the post-editing effort, i.e. the number of edits needed to correct the output hypothesis. We show that post-editing effort is very similar for top hypothesis and lower-ranked hypotheses, reinforcing the claim that lower-ranked hypotheses do not degrade in quality. Finally, we evaluate the types of edits by hypothesis rank and show that lower-ranked hypotheses propose more diverse lexical changes, in contrast to the top hypotheses that mostly attempt local spelling and grammar misuse.

Our analysis should provide insight into better training and evaluation practices for GEC. A better understanding of the under-correction phenomenon and the diversity and quality of the lower-ranked hypotheses can help improve the current training and tuning framework that relies on texts with single RGs and, arguably, hinders the development of GEC systems that can potentially address more complex linguistic phenomena.

Next, we discuss reference-based evaluation. Section 3 presents the definitions and experimental setup. Section 4 presents the evaluation using closest golds. Section 5 analyzes the edits proposed by top and lower-ranked hypotheses.

2 Reference-Based Evaluation

The standard approach to evaluating GEC systems is to use reference-based measures, that is comparing the system output to a reference that has been generated by a human annotator who corrected mistakes in the original source sentence. We refer to these as reference golds (RGs). It is common to instruct annotators to follow the principle of "minimal edits", that is making the smallest number of edits to render the sentence grammatical and well-formed. We follow a similar principle with our annotators, and the key distinction of our approach is that standard evaluations use golds that are independent of the system outputs, whereas we are creating golds by directly correcting the hypotheses output by the system. We note that there have been other proposals that argue that this principle still does not make the output fluent and propose generating references based on fluency (Sakaguchi et al., 2016). As suggested by Choshen and Abend (2018b), correcting for fluency further increases the space of valid corrections for a sentence, and we do not attempt to do this in this work.

Reference-based evaluations include several measures, such as the MaxMatch scorer M^2 (Dahlmeier and Ng, 2012), GLEU (Napoles et al., 2015), ERRANT (Bryant et al., 2017), and I-measure (Felice and Briscoe, 2015). These met-

[1] The proposed corrections are all corrections suggested by a system, some of which could be wrong.

rics have some commonalities, e.g. both MaxMatch and ERRANT measure precision, recall, and F-score. M^2 has been used with different beta parameter values, the default is $beta = 0.5$, weighting precision twice as high as recall, which is more common than assigning equal weights and has been shown to have stronger correlation with human ratings (Grundkiewicz et al., 2015). GLEU focuses on the fluency aspect – it is an extension of the BLEU metric in Machine Translation (Papineni et al., 2002). I-measure emphasizes accuracy and calculates the weighted accuracy of correction and detection. Napoles et al. (2019) proposed GMEG-Metric, that is an ensemble of existing metrics, and showed that its correlation with human judgments is higher on several GEC datasets. The MaxMatch metric has been widely used in evaluating GEC systems in many published works and in several shared tasks (Ng et al., 2013, 2014), and we adopt it in this work. We use the default beta value of 0.5 and refer to the result as F-score.

3 Definitions and Experimental Setup

We start with some definitions and then describe the experimental setup.

3.1 Definitions

Given the original learner sentence (a *source* sentence), a state-of-the-art (neural) GEC system generates a ranked list of outputs, referred to as *hypotheses*. We refer to the top hypothesis as H_1, and, similarly, to other hypotheses by the rank that they occupy. A system is evaluated using reference-based metrics where for each source sentence there is at least one corresponding corrected version that was generated by a human expert. We refer to this corrected version as *Reference Gold* (RG). The set of possible correct versions for a given source sentence is very large – possibly infinite – and any single reference gold is just a single point in that space. Most of the GEC evaluation sets contain one RG for each source sentence, although some (English) datasets contain more (CoNLL-test has 2 and an additional set of 8 generated later, and JFLEG (Napoles et al., 2017) has 4 fluency-based references). System performance is computed by scoring the top-ranked hypothesis H_1 for each sentence against the corresponding RG.

In addition to RGs, we create for each pair of (source, H_i), where H_i is the system hypothesis, another gold, which is generated by an expert by correcting the hypothesis itself. We refer to this gold as closest gold (CG_i) relative to H_i. The annotators who generated CGs were instructed to apply the minimal edit principle – i.e. correct the output to ensure it is grammatical and also preserves the meaning of the original source sentence. We thus assume that CG is as close as possible to the system output.

Given a pair of sentences, edit distance is the minimum number of edits (deletions, replacements, insertions, not necessarily single-tokens) needed, so that the sentences match. A *gold edit* is an edit between a source sentence and an RG or CG. A *proposed edit* is an edit between a source sentence and a hypothesis. A *correct edit* is an edit in the intersection of gold and proposed edits. We define Dist (S,RG) to be the number of edits between source and reference gold, and Dist (H_i,CG_i) to be the number of edits between a hypothesis H_i and CG_i relative to this hypothesis. The last one is interesting for practical purposes, since it is the post-editing effort required to completely correct the text. These are shown below.

- *S* – original text
- H_i – hypothesis at rank i
- *RG* – reference gold
- CG_i – closest gold to hypothesis H_i
- *Gold edit* – an edit between a source sentence and an RG or CG
- *Proposed edit* – an edit between a source sentence and a system hypothesis
- *Correct edit* – a proposed edit that is also a gold edit relative to a system hypothesis and specific reference
- *Dist (S,RG)* – edit distance between source and reference gold
- *Dist (H_i, RG)* – edit distance between hypothesis at rank i and reference gold
- *Dist (H_i, CG_i)* – edit distance between hypothesis at rank i and its closest gold

Table 1 shows a sample source sentence, 2 system hypotheses, the RG, and two CGs, one for each hypothesis. Dist (S,H_1) is 2 and includes 2 proposed edits ("reallistic" → "realistic" and "a" → $\varnothing$). Dist (S,RG) is 2 and includes 2 gold edits ("reallistic" → "realistic" and "had" → "gave"). The number of correct edits relative to RG and H_1 is 1 ("reallistic" → "realistic"). Dist (H_{10}, RG) is 4 (4 word replacement edits and one insertion

S	In addition , I think that the settings are very reallistic and the actors had a great performance .
$\mathbf{H_1}$	In addition , I think that the settings are very realistic and the actors had great performance .
$\mathbf{H_{10}}$	In addition , I think that the settings are very realistic and the actors performed very well .
RG	In addition , I think that the settings are very realistic and the actors gave a great performance .
$\mathbf{CG_1}$	In addition , I think that the settings are very realistic and the actors had great performances .
$\mathbf{CG_{10}}$	In addition , I think that the settings are very realistic and the actors performed very well .

Table 1: Example of an original sentence (source); the system output (hypotheses at ranks 1 and 10, H_1 and H_{10}); the reference gold (RG), and two additional golds generated on top of each of the hypotheses (CG_1 and CG_{10}).

edit), while Dist (H_{10}, CG_{10}) is 0. The three golds – two CGs and the RG – illustrate the notion of semantic equivalence (multiple ways of correcting the same source sentence, while preserving its meaning), not reflected in the standard evaluation.

3.2 Experimental Setup

We perform experiments on 2 English and 2 Russian datasets, using diverse NMT GEC model frameworks. The English datasets include the commonly used benchmarks – CoNLL-14 (Ng et al., 2014; Dahlmeier et al., 2013), and the BEA corpus (Bryant et al., 2019). The Russian datasets include the RULEC-GEC corpus (Rozovskaya and Roth, 2019) (henceforth RULEC) and another dataset of Russian learner writing that has been recently collected from the online language learning platform Lang-8 (Mizumoto et al., 2011) and annotated by native speakers.[2] We refer to this dataset as Lang8. CoNLL-14 contains two primary RGs against which the systems are standardly evaluated, while the other datasets include one RG for each sentence. We report results using one RG for each dataset for uniformity, and note that the results for the second CoNLL RG are very similar. These datasets were selected with the goal of evaluating on diverse data both in terms of genre and target language.

For the English datasets, we apply a state-of-the-art BERT-Fuse NMT system that incorporates BERT into an encoder-decoder Transformer model by (Kaneko et al., 2020). We obtained a ranked hypothesis list from the authors.

For RULEC, we use the outputs of a state-of-the-art Transformer model that uses pre-training on synthetic data and is fine-tuned on RULEC development data (Naplava and Straka, 2019). For the Lang8 corpus, we use a different state-of-the-art architecture, a Convolutional Neural Network model proposed in Chollampatt and Ng (2018b) for English. We re-implement it for Russian. The model is trained on RULEC training data and synthetic data, and uses language model re-ranking. This model is also tuned on RULEC development data. Our evaluation shows that the models are competitive: the Transformer model performs better by 4 points on the RULEC corpus than the CNN model, while the CNN model outperforms the Transformer on Lang8 by 2 points. However, we stress that our goal is not to compare these models, as we selected several model architectures and datasets to provide a more comprehensive analysis and evaluation that spans across diverse models and datasets.

Generating Closest Golds We consider the top-10 ranked hypothesis list for each dataset and study four hypotheses at the ranks 1, 2 5, and 10, to evaluate the quality of the hypotheses at various ranks and to determine how much quality degrades from the top hypothesis downwards. For each of the 4 hypotheses H_i, $i \in \{1, 2, 5, 10\}$, a closest gold CG_i relative to this hypothesis is generated by post-editing the hypothesis for grammatical errors and other misuse.

Annotation 100 source sentences from each dataset were selected uniformly at random and 4 hypotheses at different ranks were annotated for each sentence. The English outputs were annotated by two annotators – one native English speaker and a fluent non-native speaker. Each annotator corrected all hypotheses for one of the datasets. This was done to ensure consistency across different hypotheses. The Russian outputs were corrected by one native Russian speaker. All of the annotators have a Master's degree and previous annotation experience. The annotators followed the standard annotation protocol in grammar correction, in that they were instructed to follow the minimal-edits principle in correcting the sentences, while also ensuring the output is well-formed and adequate (i.e. the meaning of the original source sentence is preserved), for which they also consulted the source sentence.

[2] This is a recently collected dataset that will be made available for research.

4 Evaluating True System Performance

We start by evaluating each hypothesis output H_i for each dataset against reference gold RG and its corresponding closest gold CG_i. We show that evaluation relative to RG is always pessimistic and, given a hypothesis generated by a GEC system, there is always a much better gold.

4.1 Reference Gold vs. Closest Gold

Table 2 shows the results of evaluating each system hypothesis against reference golds and closest golds for two datasets – BEA (English) and RULEC (Russian). Results for all datasets are in Appendix (Table 7). The CG result is significantly higher than the performance relative to RG in all cases. For the top hypothesis, the F-scores increase by 19 points on BEA and 17 points on RULEC. Improvements are greater for lower-ranked hypotheses. The improvements for BEA are 34, 36, and 37, for ranks 2, 5 and 10, respectively, and for RULEC – 34, 28, and 31.

The most substantial changes occur in precision: between 23 and 41 points on BEA, and 24 and 40 for RULEC (similar changes for the other datasets). It should be emphasized that precision improvements relative to RG are greater for lower-ranked hypotheses. This is interesting and suggests that while lower-ranked hypotheses propose significantly more changes than the top-ranked one (see column “Proposed edits”), a lot of those edits are valid corrections, even though they are not recognized in RGs: observe that despite the fact that more edits are being proposed with lower hypotheses rank, the number of correct edits (shown in the table) relative to RG goes down. For instance, 84 out of 125 proposed edits are correct in BEA H_1, while only 75 out of 200 proposed are valid in H_5. This is consistent across the datasets and indicates that *changes proposed in lower-ranked hypotheses are less likely to be included in the RGs*.

Recall is also improved in CGs relative to RGs, although not as dramatically. Recall increases by 10-25 points on CoNLL, 12-34 points for BEA, 7-22 points for RULEC, and 2-12 points for Lang8.

The results in the table strongly indicate that the n-best list does not produce hypotheses of degrading quality. On the contrary, the precision of the proposed corrections remains impressively high (in most cases, well above 50 and often into 70 or 80), which is not reflected in the reference-based evaluation scoring against a reference gold. We further illustrate the finding in Figure 2, where for each dataset, we show F-scores of the 4 hypotheses against RGs, and scores against their corresponding CGs. The first observation is that performances in the first group are much lower than in the second group for each corpus. But it is also clear that the first group shows strong degradation relative to the top-ranked hypothesis in the RG evaluation – performance goes down as you go from H_1 to H_{10}, while in the second group the performance remains almost the same across the four hypotheses.

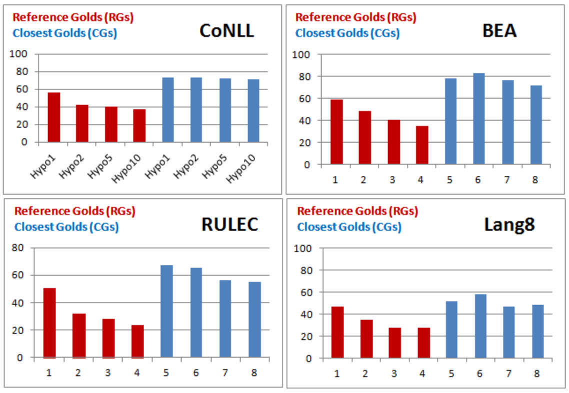

Figure 2: Performance by hypothesis rank (F-score) against Reference Gold (RG) vs. Closest Golds (CGs) generated specially for each hypothesis. Observe dramatic drop in performance between top hypothesis and the rest in RG evaluation, vs. stability in CG evaluation; and large gaps between scores even for the top hypotheses in RG vs. CG evaluations.

Further, as shown in Table 2, the number of correct edits is significantly higher when evaluated against CGs. For instance, the number of correct edits increases from 75 to 163 for BEA H_5, and from 48 to 105 for RULEC H_2. Additionally, the number of gold edits in CGs is much higher than in RGs, and is also greater for lower-ranked hypotheses. For instance, there are 202 golds edits in BEA RG, 217 edits in BEA CG_1 and 282 edits in BEA CG_{10}. This is consistent across the datasets and suggests that *the edits proposed by the models are not necessarily the minimal edits that most of the GEC annotations adopt*. This may be why most of the proposed edits in the lower-ranked hypotheses are not found in the RGs. Table 8 in Appendix also evaluates each hypothesis against CGs relative to the other hypotheses, showing that evaluation against CG always produces superior results.

Dataset	Hypo	Gold type	P	R	F-score	Edits		
						Correct	Proposed	Gold
BEA	H_1	RG	65.9	40.1	58.4	84	125	202
		CG_1	88.4	52.5	**77.8**	114	125	217
	H_2	RG	50.0	43.1	48.4	87	180	202
		CG_2	87.8	66.4	**82.5**	158	180	238
	H_5	RG	41.2	37.1	40.3	75	200	202
		CG_5	81.5	61.0	**76.4**	163	200	267
	H_{10}	RG	34.8	35.1	34.9	71	220	202
		CG_{10}	75.0	59.6	**71.3**	168	220	282
RULEC	H_1	RG	63.6	27.7	50.5	56	90	202
		CG_1	87.5	34.7	**67.1**	77	90	222
	H_2	RG	34.8	23.8	31.8	48	144	202
		CG_2	74.5	43.4	**65.1**	105	144	242
	H_5	RG	29.2	24.3	28.0	49	174	202
		CG-H_5	61.6	41.6	**56.2**	109	174	262
	H_{10}	RG	24.0	21.3	23.4	43	194	202
		CG_{10}	59.1	43.0	**55.0**	110	194	256

Table 2: Performance by hypothesis rank against reference gold (RG) and Closest Golds (CGs) generated specially for each hypothesis. For each hypothesis, number of correct, proposed, and gold edits relative to each reference (RG or CG) are also shown. Results for all the datasets are in Appendix (Table 7).

4.2 Quality Estimation with Edit Rate

We have shown above that the quality of the hypotheses does not degrade with rank, and in some cases hypotheses at lower ranks even result in higher F-score than the top-ranked hypothesis, when scored against CG, while the evaluation against RGs is strongly biased against lower-ranked hypotheses. We now wish to evaluate hypotheses quality using the edit rate, i.e. the number of edits needed to fix the output hypothesis so that it matches its corresponding CG. This quality estimation approach that considers the number of edits required to "fix" the hypothesis is used in Machine Translation (Snover et al., 2006), where the quality of a system output is measured as the minimum number of edits needed to transform the system output so that it matches a reference. To this end, a "targeted" reference is created for a translated sentence, by editing the hypothesis until it is both fluent and has the same meaning as the (original) reference(s). The reason for this is that estimating quality against gold "non-targeted" reference ignores notions of semantic equivalence (see also Table 1), thereby underestimating output quality. Thus, a targeted reference provides a more accurate measure of translation quality. Chollampatt and Ng (2018a) proposed a quality estimation model for GEC that builds on this idea of measuring output quality as the number of edits required to fix the hypothesis. However, they make the strong assumption that, unlike in MT, in GEC targeted references need not be created, as RGs can be substituted for CGs, because both human annotators and automatic GEC systems are trained to make minimal changes. As we showed in the previous section, this is not the case and *using RGs severely underestimates system performance, and, as a consequence, post-editing effort.*

We now use CGs to estimate hypotheses quality in terms of post-editing effort in Table 3. We show the number of proposed edits, the number of correct edits relative to CG, and the number of gold edits in the corresponding CG. (The number of proposed, gold, and correct edits also appears in Tables 2 and 7 but is shown here in Table 3 for convenience). The post-editing effort is shown in the last column, estimated as the number of edits required to make the hypothesis output fluent and grammatical, i.e. the edit distance between a hypothesis and its corresponding CG. The number of edits was computed automatically using the ERRANT tool (Bryant et al., 2017) that, given a pair of sentences (source, hypothesis), will produce a set of edits needed to transform the source into the target. The post-editing effort is not necessarily the smallest for the top hypothesis. In fact, on BEA, the smallest value is obtained for H_2 (55 edits), while H_1 and H_{10} are similar (86 and 87). On the other datasets, there is no significant difference for the English datasets across the 4 different hypotheses, while on the Russian datasets there is slight degradation for hypotheses 5 and 10, while H_1 and H_2 are close. This supports our finding above that lower-ranked hypotheses are of high quality. As a side note, our post-edit estimation assumes that there are no errors that have

Dataset	Hypo	Edits			
		Prop.	Corr.	Gold	Post-Edit
CoNLL	H_1	156	132	300	**130**
	H_2	203	160	349	159
	H_5	239	184	343	138
	H_{10}	266	196	349	142
BEA	H_1	125	114	217	86
	H_2	180	158	238	**55**
	H_5	200	163	267	73
	H_{10}	220	168	282	87
RULEC	H_1	90	88	222	**119**
	H_2	144	105	242	131
	H_5	174	109	262	173
	H_{10}	194	110	256	186
Lang8	H_1	98	65	252	**168**
	H_2	186	117	287	174
	H_5	214	105	298	215
	H_{10}	225	109	312	225

Table 3: Number of edits by hypothesis rank. We show the number of proposed edits, the number of correct edits relative to CG, the number of gold edits in CG, and the post-editing effort required to make the hypothesis fluent and grammatical, estimated as the number of edits between the hypothesis and its CG.

more impact than others. In Section 5, we actually show that the top-ranked hypotheses mostly contain changes on “simpler errors”, and, arguably, the lower-ranked hypotheses might even involve less post-editing effort given the more complex errors they manage to fix.

4.3 Do GEC Systems Undercorrect?

We first compare the number of edits in each hypothesis to the number of edits in the original gold (Table 4). The top-ranked hypothesis makes only a fraction of edits compared to RGs. Generally, RGs contain 2.5-3 more edits than the top-1 hypothesis. This is consistent with the analysis in Choshen and Abend (2018b) that shows that GEC systems are disincentivized to make corrections due to the low-coverage bias. What is notable, however, is that the number of edits substantially increases with the hypothesis rank. In particular, *the second-ranked hypothesis contains on average twice as many edits as the first one*, and the number of edits continues to increase. Hypotheses 5 and 10 contain a similar number of edits compared to the number of edits in RG. The under-correction issue is further studied in the next section.

Hypo	Number of edits proposed			
	RULEC	Lang8	BEA	CoNLL
	(2,646)	(2,260)	(2,103)	(2,665)
H_1	90	98	125	156
H_2	144	186	180	203
H_5	174	214	200	239
H_{10}	194	225	220	266
RG	202	232	202	289

Table 4: Number of proposed edits by hypothesis rank. For each corpus, total number of tokens is shown. The majority of edits are single-token replacements, deletions or insertions. The last row shows the number of gold edits in the *reference gold* for each dataset.

5 Edit Analysis by Hypothesis Rank

We now analyze and compare the edits in the top-ranked hypothesis and in H_{10}, in order to understand better how the edits differ with hypothesis rank. For the English datasets, we apply ERRANT (Bryant et al., 2017) to extract edits using pairs of parallel sentences (source, hypothesis). ERRANT then uses English-language specific rules based on part-of-speech and linguistic knowledge to assign each edit its linguistic type, such as preposition, noun number, etc. We further group the edits into one of the following two categories: *spelling/grammar changes* and *lexical changes*. The first category includes punctuation, spelling, orthography, and grammatical corrections that typically require local context and small changes and are also limited in the number of candidate corrections. These include determiner errors, verb agreement and form, noun number and punctuation, and morphological changes. Lexical changes comprise the categories denoted by ERRANT as “Other”, “Verb”, “Noun”, “Pronoun”, “Adverb”, which include mostly lexical errors, e.g. changing “get” to “earn”, verb tense errors that require wider context and thus are trickier to correct. The number of edits by type is shown in Table 5. Lexical changes are marked with a (*).

In the lower part of the table, we show the distribution of edits between the two categories: in CoNLL, spell/grammar changes account for 51.2% of all changes in the RG and for 74.3% in the top-ranked hypothesis. Lexical changes make up 48.8% in RG, while only 25.7% in the top-ranked hypothesis, although this number increases to 36.2% in the H_{10}. In BEA, 49.5% of RG edits are lexical, while in the top-ranked hypothesis

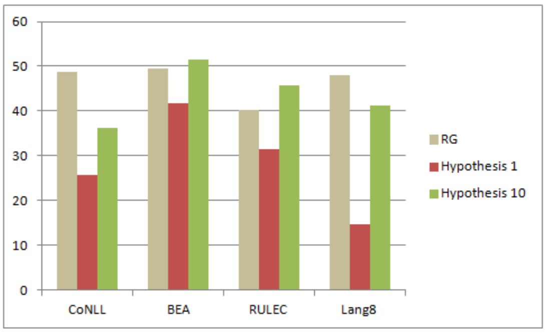

Figure 3: Percentage of lexical edits with respect to total number of changes in reference gold, H_1, and H_{10}.

these account for 41.7% and that number goes up to 51.5% for the H_{10}.

Looking at the number of edits in each category, it can be observed, that the under-correction phenomenon (for the top-ranked hypothesis) is particularly pronounced for lexical errors. In the spell-grammar category, the number of proposed edits is very close to (or even exceeds) the number of edits of this type in RG in both datasets (the only exception is perhaps the punctuation errors). For example, 33 determiner errors are present in CoNLL RG, while there are 25 in H_1. In contrast, 37 errors of category "Other" are in RG for CoNLL but only 8 changes of this type are in the top-ranked hypothesis. In fact, in both CoNLL and BEA, in the top-ranked hypotheses, the majority of the changes are minimal/local changes (74.3% in the CoNLL dataset and 58.3% in the BEA dataset).

We perform a similar analysis for the Russian datasets, where the edits are classified manually by our annotator (due to lack of automatic tool). We find similar behavior (see Appendix A). However, in the most challenging categories (lexical and "Other"), which both comprise word changes, the situation is more severe: the top-ranked hypothesis proposes 0 changes. Overall, the under-correction phenomenon for lexical errors is more pronounced for the Russian language.

Overall, the under-correction phenomenon is especially pronounced for top hypotheses in the lexical error category. The percentage of lexical edits with respect to the total number of edits in the RG and CGs is much higher than in the top hypotheses. Thus, under-correction is mostly a problem for lexical errors, but is partially rectified in the lower-ranked hypotheses, illustrated in Figure 3 that shows the percentage of lexical edits in RG, H_1, and H_{10} for each dataset.

Edit type	**CoNLL**			**BEA**		
	RG	H_1	H_{10}	RG	H_1	H_{10}
Spell/Orth	10	21	23	16	12	14
Punc	9	3	19	35	12	30
Noun number	19	25	35	7	8	14
Det	33	25	52	22	21	28
Verb agr.	14	15	13	3	3	2
Verb form	15	13	10	7	5	6
Morph.	6	5	5	5	6	6
Prep*	18	14	21	22	9	24
Verb tense*	18	7	17	11	8	19
Other*	37	8	25	37	19	38
Verb*	12	3	9	10	5	6
Noun*	4	1	9	9	2	6
Pronoun*	6	3	6	4	2	3
Adverb*	5	0	0	3	3	10
Spell/grammar (%)	51.2	74.3	63.8	50.5	58.3	48.5
Lex. changes (%)	48.8	25.7	36.2	49.5	41.7	51.5

Table 5: Proposed edits and gold RG edits by type and hypothesis rank on the English datasets. * marks lexical changes.

6 Discussion

We study the current evaluation and training schema in GEC, using 4 datasets in 2 languages and several state-of-the-art model architectures. We make several observations. First, we show that the quality of the systems is significantly better than we think, when we evaluate relative to the closest gold vs. reference gold. And the reason is there are many golds and we show that there is always a gold that is close to the prediction, and we should take this result as the actual performance of the model.[3] Moreover, as we showed, using the CGs provides additional knowledge about the type of errors various hypotheses generate, further guiding the community towards developing additional insights that can be used also in targeting specific models for specific users (based on their abilities, for example). Our second observation is that the top hypothesis is not actually better than the lower-ranked hypotheses in the 10-best list, even though the current evaluations are strongly biased towards the top hypothesis. Third, because

[3] In fact, given that sentence-level ensembles computed via multiple references were shown to perform much better than a single hypothesis (Bryant and Ng, 2015), already indicates the existence of better golds, since a sentence-based combination of reference gold is a gold by itself. We claim, though, that even this underestimates the true performance, as shown in our evaluation relative to the CGs.

of the way we train, lower-ranked hypotheses relative to the reference gold are as good or sometimes qualitatively better than the top hypothesis because of the diversity of the type of mistakes that they attempt to correct.

Recommendations based on the paper findings We view this paper as an analysis paper that we hope can contribute to a better understanding of the current issues in the GEC field. We hope that the proposed analysis can give an opportunity to researchers to think about directions for addressing these issues. That said, we believe that our results may serve as a preliminary proposal for developing better ways for evaluating for GEC systems, and would like to outline several recommendations based on our findings. We believe the findings should be useful for thinking about how to modify the training and tuning paradigm in GEC.

Regarding training and tuning, the current schema of using learner texts with single RGs hinders development of GEC systems that, as we show, can potentially address more complex linguistic phenomena and language misuse. For training and tuning, perhaps, it would make sense to generate multiple references by creating additional references that contain paraphrases of the original gold reference. In terms of evaluation, the findings might inspire researchers to think of better ways to evaluate GEC system outputs. For example, instead of computing exact match, we could include paraphrases so as not to penalize hypotheses that propose more liberal sentence-rewrites. A different approach might be to choose lower-ranked hypotheses, since they are as good, and they have some other useful properties, such as the language phenomena they are able to correct that the top hypothesis cannot.

Acknowledgments

We thank the annotators Mikhail Khabas and Maxwell White. We thank the reviewers for their comments. This work was supported by Contract FA8750-19-2-0201 with the US Defense Advanced Research Projects Agency (DARPA).

References

C. Bryant, M. Felice, Ø. Andersen, and T. Briscoe. 2019. The BEA-19 shared task on grammatical error correction. In *Proceedings of the ACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA-19)*.

C. Bryant, M. Felice, and T. Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *ACL*.

C. Bryant and H. T. Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In *ACL*.

M. Chodorow, M. Dickinson, R. Israel, and J. Tetreault. 2012. Problems in evaluating grammatical error detection systems. In *Proceedings of COLING*.

S. Chollampatt and H.T. Ng. 2018a. Neural quality estimation of grammatical error correction. In *EMNLP*.

Shamil Chollampatt and Hwee Tou Ng. 2018b. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the AAAI*. Association for the Advancement of Artificial Intelligence.

L. Choshen and O. Abend. 2018a. Automatic metric validation for grammatical error correction. In *ACL*.

L. Choshen and O. Abend. 2018b. Inherent biases in reference-based evaluation for grammatical error correction and text simplification. In *ACL*.

D. Dahlmeier and H. T. Ng. 2012. A beam-search decoder for grammatical error correction. In *Proceedings of EMNLP-CoNLL*.

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, Georgia. Association for Computational Linguistics.

M. Felice and T. Briscoe. 2015. Towards a standard evaluation method for grammatical error detection and correction. In *NAACL-HLT*.

R. Grundkiewicz, M. Junczys-Dowmunt, and E. Gillian. 2015. Human evaluation of grammatical error correction systems. In *EMNLP*.

M. Kaneko, M. Mita, S. Kiyono, J. Suzuki, and K. Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *ACL*.

T. Mizumoto, M. Komachi, M. Nagata, and Y. Matsumoto. 2011. Mining revision log of language learning SNS for automated japanese error correction of second language learners. In *IJCNLP*.

J. Naplava and M. Straka. 2019. Grammatical error correction in low-resource scenarios. In *W-NUT Workshop*.

C. Napoles, M. Nădejde, and J. Tetreault. 2019. Enabling robust grammatical error correction in new domains: Data sets, metrics, and analyses. In *Transactions of ACL*.

C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault. 2015. Ground truth for grammatical error correction metrics. In *ACL*.

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

H. T. Ng, S. M. Wu, Y. Wu, Ch. Hadiwinoto, and J. Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of CoNLL: Shared Task*.

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

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

A. Rozovskaya and D. Roth. 2019. Grammar error correction in morphologically-rich languages: The case of russian. In *Transactions of ACL*.

K. Sakaguchi, C. Napoles, M. Post, and J. Tetreault. 2016. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. In *Transactions of ACL*.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Association for Machine Translation in the Americas*.

A Edit Analysis for the Russian Datasets

For the Russian corpora, since there is no automated tool that classifies edits by type, we manually classify the edits in the hypotheses and RGs. We first extract all proposed edits using the ERRANT tool (The errant tool both extracts edits given a pair of sentences, and classifies these by type). The edit-extraction component is language-independent, whereas the type-classification is English-based). These edits are then manually classified into one of the grammar categories relevant for Russian. We use the error classification schema in Rozovskaya and Roth (2019) but combine certain types, e.g. we group together noun and adjective case errors, verb tense/aspect errors, and noun and adjective number errors. Unlike English, Russian does not have determiner errors. We similarly group the error type into two categories, spelling/grammar and lexical. However, we assign the noun/adjective number errors and morphology errors to the second group (lexical), as we expect these to display more variability due to the number of of different endings for adjective/noun number because of declensions and gender, and the large number of morphological variants compared to English. The statistics are shown in Table 6.

First, observe that the distribution in the gold references of lexical and spelling/grammar changes in the lower part of the table is similar to the English datasets. 40% or more of all gold edits are lexical. In the top-ranked hypothesis, only 31.5% and 14.6% of edits in RULEC and Lang8, respectively are of this type. This is similar to the results for the English datasets in Table 5, however in the most challenging categories (lexical and "other"), which both comprise word changes, the situation is more severe: the top-ranked hypothesis proposes 0 changes. Overall, the under-correction phenomenon is more pronounced for the Russian language.

Edit	**RULEC**			**Lang8**		
type	RG	H_1	H_{10}	RG	H_1	H_{10}
Spell	40	29	35	53	44	54
Noun/adj. case,gender,num.	46	20	43	39	19	29
Verb agr.	5	1	4	6	5	4
Punc	24	17	32	16	11	34
Prep*	11	6	18	14	6	23
Verb tense/aspect*	6	3	20	12	1	20
Noun/adj. num.*	7	1	17	13	4	23
Morph.*	15	5	13	19	4	15
Lexical*	30	0	13	30	0	5
Other changes*	20	0	4	32	0	12
Spell/grammar (%)	59.7	68.5	54.2	52.1	85.4	58.8
Lex.changes (%)	40.3	31.5	45.8	47.9	14.6	41.2

Table 6: Proposed edits and gold RG edits by type and hypothesis rank on the Russian datasets. * marks lexical changes.

B Additional Results for All Datasets

Dataset	Hypo	Gold type	P	R	F-score	Proposed edits		
						Correct	Proposed	Gold
CoNLL	H_1	RG	66.7	33.9	55.9	98	156	289
		CG_1	87.4	44.0	**73.0**	132	156	300
	H_2	RG	48.0	29.8	42.8	86	203	289
		CG_2	87.4	44.0	**73.0**	160	203	349
	H_5	RG	43.2	31.8	40.3	92	239	289
		CG_5	79.7	53.6	**72.6**	184	239	343
	H_{10}	RG	39.6	31.1	37.6	90	266	289
		CG_{10}	76.6	56.2	**71.4**	196	266	349
BEA	H_1	RG	65.9	40.1	58.4	84	125	202
		CG_1	88.4	52.5	**77.8**	114	125	217
	H_2	RG	50.0	43.1	48.4	87	180	202
		CG_2	87.8	66.4	**82.5**	158	180	238
	H_5	RG	41.2	37.1	40.3	75	200	202
		CG_5	81.5	61.0	**76.4**	163	200	267
	H_{10}	RG	34.8	35.1	34.9	71	220	202
		CG_{10}	75.0	59.6	**71.3**	168	220	282
RULEC	H_1	RG	63.6	27.7	50.5	56	90	202
		CG_1	87.5	34.7	**67.1**	77	90	222
	H_2	RG	34.8	23.8	31.8	48	144	202
		CG_2	74.5	43.4	**65.1**	105	144	242
	H_5	RG	29.2	24.3	28.0	49	174	202
		CG_5	61.6	41.6	**56.2**	109	174	262
	H_{10}	RG	24.0	21.3	23.4	43	194	202
		CG_{10}	59.1	43.0	**55.0**	110	194	256
Lang8	H_1	RG	60.9	24.1	46.7	56	98	232
		CG_1	69.2	25.8	**51.8**	65	98	252
	H_2	RG	36.9	28.5	34.8	66	186	232
		CG_2	64.6	40.8	**57.9**	117	186	287
	H_5	RG	28.6	24.1	27.6	56	214	232
		CG_5	50.7	34.9	**46.5**	105	214	298
	H_{10}	RG	28.4	23.7	27.3	55	225	232
		CG_{10}	53.2	34.9	**48.1**	109	225	312

Table 7: Performance by hypothesis rank against reference gold (RG) and Closest Golds (CGs) generated specially for each hypothesis. For each hypothesis, number of correct, proposed, and gold edits relative to each gold are also shown. Expanded version of Table 2 in Section 4.1 that showed results for BEA and RULEC only.

Dataset	Hypo	Gold type	P	R	F-score
CoNLL	H_1	RG	66.7	33.9	55.9
		Other CGs	74.0-75.8	31.8-32.9	58.5-60.2
		CG_1	87.4	44.0	**73.0**
CoNLL	H_2	RG	48.0	29.8	42.8
		Other CGs	58.9-59.6	31.2-37.3	50.0-53.2
		CG_2	83.8	45.8	**71.9**
CoNLL	H_5	RG	43.2	31.8	40.3
		Other CGs	51.8-52.3	33.0-38.7	46.5-48.8
		CG_5	79.7	53.6	**72.6**
CoNLL	H_5	RG	39.6	31.1	37.6
		Other CGs	48.8-50.2	33.8-41.0	45.2-48.0
		CG_{10}	76.6	56.2	**71.4**
BEA	H_1	RG	65.9	40.1	58.4
		Other CGs	74.2-75.2	33.3-39.9	60.1-63.7
		CG_1	88.4	52.5	**77.8**
BEA	H_2	RG	50.0	43.1	48.4
		Other CGs	60.3-67.2	41.9-49.8	56.7-60.1
		CG_2	87.8	66.4	**82.5**
BEA	H_5	RG	41.2	37.1	40.3
		Other CGs	48.7-52.7	35.1-42.4	47.3-48.6
		CG_5	81.5	61.0	**76.4**
BEA	H_{10}	RG	34.8	35.1	34.9
		Other CGs	44.5-47.7	38.2-44.7	44.5-45.7
		CG_{10}	75.0	59.6	**71.3**
RULEC	H_1	RG	63.6	27.7	50.5
		Other CGs	79.1-80.7	27.1-28.1	57.8-58.0
		CG_1	87.5	34.7	**67.1**
RULEC	H_2	RG	34.8	23.8	31.8
		Other CGs	49.3-55.3	26.3-31.1	42.0-47.6
		CG_2	74.5	43.4	**65.1**
RULEC	H_5	RG	29.2	24.3	28.0
		Other CGs	36.3-38.0	24.6-29.3	33.4-35.9
		CG_5	61.6	41.6	**56.2**
RULEC	H_{10}	RG	24.0	21.3	23.4
		Other CGs	32.8-35.6	24.0-27.0	31.4-33.3
		CG_{10}	59.1	43.0	**55.0**
Lang8	H_1	RG	60.9	24.1	46.7
		Other CGs	64.8-69.6	18.9-21.5	43.9-48.1
		CG_1	69.2	25.8	**51.8**
Lang8	H_2	RG	36.9	28.5	34.8
		Other CGs	39.9-45.6	26.3-28.2	36.8-39.7
		CG_2	64.6	40.8	**57.9**
Lang8	H_5	RG	28.6	24.1	27.6
		Other CGs	32.8-34.4	21.5-25.4	30.6-31.0
		CG_5	50.7	34.9	**46.5**
Lang8	H_{10}	RG	28.4	23.7	27.3
		Other CGs	29.2-33.9	21.8-22.2	27.5-30.5
		CG_{10}	53.2	34.9	**48.1**

Table 8: Performance by hypothesis rank against reference gold (RG) and Closest Golds (CGs) generated specially for each hypothesis.