Revisiting Meta-evaluation for Grammatical Error Correction

Masamune Kobayashi[◇] **Masato Mita**[†◇] **Mamoru Komachi**[‡]
[◇]Tokyo Metropolitan University, Japan [†]CyberAgent Inc., Japan [‡]Hitotsubashi University, Japan
kobayashi-masamune@ed.tmu.ac.jp, mita_masato@cyberagent.co.jp,
mamoru.komachi@r.hit-u.ac.jp

Abstract

Metrics are the foundation for automatic evaluation in grammatical error correction (GEC), with their evaluation of the metrics (meta-evaluation) relying on their correlation with human judgments. However, conventional meta-evaluations in English GEC encounter several challenges, including biases caused by inconsistencies in evaluation granularity and an outdated setup using classical systems. These problems can lead to misinterpretation of metrics and potentially hinder the applicability of GEC techniques. To address these issues, this paper proposes SEEDA, a new dataset for GEC meta-evaluation. SEEDA consists of corrections with human ratings along two different granularities: *edit-based* and *sentence-based*, covering 12 state-of-the-art systems including large language models, and two human corrections with different focuses. The results of improved correlations by aligning the granularity in the sentence-level meta-evaluation suggest that edit-based metrics may have been underestimated in existing studies. Furthermore, correlations of most metrics decrease when changing from classical to neural systems, indicating that traditional metrics are relatively poor at evaluating fluently corrected sentences with many edits.

1 Introduction

Grammatical error correction (GEC) is the task of automatically detecting and correcting errors, including grammatical, orthographic, and semantic errors, within a given sentence. The prevailing approach in GEC involves the use of a sequence-to-sequence method (Bryant et al., 2023).

Automatic evaluation metrics play an important role in the progress of GEC. These metrics are essential for a fast and efficient improvement cycle of system development because they can replace costly and time-consuming human evaluations and immediately reflect system performance. GEC has made progress by enabling a fair comparison of performance on a common benchmark using these metrics in shared tasks (Dale and Kilgarriff, 2011; Dale et al., 2012; Ng et al., 2014; Bryant et al., 2019).

GEC metrics are categorized into *edit-based* and *sentence-based* types according to their evaluation granularity, and each has its objectives. Edit-Based Metrics (EBMs), such as M^2 (Dahlmeier and Ng, 2012) and ERRANT (Bryant et al., 2017), focus on evaluating the quality of the edit itself, whereas Sentence-Based Metrics (SBMs), such as GLEU (Napoles et al., 2015), evaluate the quality of the entire sentence after correction. Since the system output consists only of sentences without explicit edits, EBMs require the edit extraction from the system output using any method. In addition, these metrics are primarily evaluated based on the correlation with human judgment (i.e., *meta-evaluation*).

Most of the previous meta-evaluations in English GEC have relied on Grundkiewicz et al.'s (2015) dataset with human judgments (henceforth, this dataset is referred to as *GJG15*). However, existing meta-evaluations based on GJG15 (Grundkiewicz et al., 2015; Chollampatt and Ng, 2018b; Yoshimura et al., 2020; Gong et al., 2022) have several significant issues. First, the performance of EBMs may be underestimated due to biases resulting from inconsistencies in evaluation granularity. As an example of biases, while EBMs assign the lowest score (or the highest score in the sentence-level evaluation) to the uncorrected sentence, sentence-based human evaluation, such as GJG15, assigns scores across the entire range. Furthermore, according to the actual data in Table 1, since human evaluations may yield different results based on granularity, the GEC evaluation suggests a need to separate evaluations for edits and sentences. Second, GJG15 is manually evaluated against the set of classical systems in the CoNLL-2014 shared task, such as

Transactions of the Association for Computational Linguistics, vol. 12, pp. 837–855, 2024. https://doi.org/10.1162/tacl_a_00676
Action Editor: Nathan Schneider. Submission batch: 11/2023; Revision batch: 7/2024; Published 7/2024.

	Grammatical Error Correction
Source:	It is hereditary. Do one who suffered from this disease keep it a secret of infrom their relatives? In retrospect, its is also ones duty to ensure that he or she undergo periodic healthchecks in their own.
Output A:	Should someone who suffered from this disease keep it a secret or inform their relatives?
Output B:	Does someone who suffers from this disease keep it a secret from their relatives?
	Edit-based human evaluation
Output A:	***(Rank 1)*** [Do → Should] [one → someone] who suffered from this disease keep it a secret [of → or] [inform → inform] their relatives?
Output B:	***(Rank 1)*** [Do → Does] [one → someone] who [suffered → suffers] from this disease keep it a secret [of inform → from] their relatives?
	Sentence-based human evaluation
Output A:	***(Rank 1)*** Should someone who suffered from this disease keep it a secret or inform their relatives?
Output B:	***(Rank 5)*** Does someone who suffers from this disease keep it a secret from their relatives?

Table 1: Actual data taken from our dataset shows that the results of human evaluation vary depending on the granularity. In edit-based evaluation, output B was assigned the *highest* rank (tied with output A), while in sentence-based evaluation, output B received the *lowest* rank. The results suggest that, even if all edits are considered valid, there are instances where the corrected sentence may lack fluency and naturalness in context.

statistical machine translation approach (Junczys Dowmunt and Grundkiewicz, 2014), and classifier-based approach (Rozovskaya et al., 2014). Therefore, the gap between the classical systems in GJG15 and the current modern GEC systems based on deep neural networks limits the applicability of meta-evaluation. Third, a single correlation from the current fixed set of systems may not sufficiently capture the performance of metrics, leading to the possibility of drawing incorrect conclusions. For example, Deutsch et al.'s (2021) study on meta-evaluation of summarization revealed that certain metrics can exhibit a spectrum of correlation values, ranging from weak negative to strong positive correlations. Mathur et al.'s (2020) study also showed that outlier systems have a strong influence on correlations in a meta-evaluation of machine translation. Therefore, we are concerned that a similar scenario could occur in the GEC.

To address these issues, we propose SEEDA,[1] a new dataset to improve the validity of meta-evaluation in English GEC. Specifically, we carefully designed SEEDA to address the first and second issues by performing human evaluations corresponding to two different granularity metrics (i.e., EBMs and SBMs), covering 12 state-of-the-art system corrections including large language models (LLMs), and two human corrections with different focuses (§3 and §4). Also, through meta-evaluation using SEEDA, we investigate whether EBMs, such as M^2 and ERRANT, are underestimated and demonstrate how the correlation varies between classical systems and neural systems (§6). Furthermore, to address the third issue, we investigate the inadequacy of GEC meta-evaluation based solely on a single correlation by analyzing the presence of outliers and using window analysis (§7). Finally, we discuss best practices and provide recommendations for future researchers to properly meta-evaluate GEC metrics and evaluate their GEC models (§8).

Our contributions are summarized as follows. (1) We construct a new dataset that allows for bias-free meta-evaluation that fits modern neural systems. (2) The dataset analysis shows variations in sentence-level human evaluation results depending on the evaluation granularity. (3) We

[1]SEEDA stands for **S**entence-based and **E**dit-based human **E**valuation **DA**taset for GEC. We have made this dataset publicly available at `https://github.com/tmu-nlp/SEEDA`.

identified two findings through meta-evaluation: aligning the granularity between human evaluation and metric enhances correlations, and correlations for classical and neural systems are different. (4) Investigating the influence of outliers and system sets, we discovered that a meta-evaluation of a single setting cannot analyze the detailed characteristics of the metric. We also found that existing metrics lack the precision to differentiate between the performances of top-tier systems.

2 Related Work

Meta-evaluation Grundkiewicz et al. (2015) proposed a dataset (GJG15) with sentence-based human ratings for system outputs in the CoNLL-2014 test set and found that M^2 has a moderate positive correlation with human judgments. Simultaneously, Napoles et al. (2015) constructed a dataset by performing a similar human evaluation and observed that their proposed metric, GLEU, has a stronger correlation than M^2. Both studies found no correlation with I-measure (Felice and Briscoe, 2015). Chollampatt and Ng (2018b) carried out significance tests between various metrics using GJG15. They concluded that there was no clear distinction in performance between M^2 and GLEU, with I-measure proving to be the most robust metric. However, these experiments are based on classical systems and thus deviate from modern neural systems. MAEGE proposed by Choshen and Abend (2018a) applies multiple partial edits to the uncorrected sentence and assigns pseudo-scores based on the number of edits, aiming for a meta-evaluation independent of human evaluation. MAEGE does not consider system outputs and human evaluations, so it should be distinguished from existing meta-evaluations that rely on humans. Moreover, since it does not account for errors that machines might make but humans wouldn't, the need for human evaluation against outputs persists. Furthermore, Napoles et al. (2019) constructed GMEG-Data by performing human judgments using continuous scales on the CoNLL-2014 test set and three domain-specific datasets. Their findings highlighted diverse correlations across the different domains. They explored neural systems, but these deviate from mainstream systems pretrained with pseudo data and fine-tuned based on the transformer (Vaswani et al., 2017). While SEEDA offers greater validity due to its focus on contemporary target systems and the evaluation granularity, GMEG-Data has the advantage of allowing meta-evaluation using the entire CoNLL-2014 benchmark in various domains.

Reference-based Evaluation In the evaluation of GEC, commonly used metrics rely on reference sentences. Some of the most prevalent metrics include M^2, ERRANT, and GLEU. Both M^2 and ERRANT calculate $F_{0.5}$ score by comparing the edits in the corrected sentence to those in the reference. In contrast, GLEU assesses based on the matching of N-grams between the corrected sentence and reference. I-measure evaluates the degree of improvement from the original sentence using the weighted precision of edits. There are also newer metrics like GoToScorer (Gotou et al., 2020), which takes into account the difficulty of corrections, and PT-M^2 (Gong et al., 2022), which extends M^2 (and ERRANT) with pretraining-based metrics. It is worth noting that these reference-based evaluations can lose validity with limited reference coverage.

Reference-less Evaluation Evaluations without reference sentences aim to overcome the coverage issues. GBM (Napoles et al., 2016b) estimates grammaticality by identifying the number of errors in a sentence. However, it may be less sensitive to semantic changes. To address this limitation, GFM (Asano et al., 2017) was proposed. It incorporates sub-metrics to estimate grammaticality, fluency, and meaning preservation. Additionally, USim (Choshen and Abend, 2018b) was developed to specifically estimate semantic faithfulness. SOME (Yoshimura et al., 2020) draws inspiration from GFM and optimizes each sub-metric based on human evaluation using BERT (Devlin et al., 2019). Scribendi Score (Islam and Magnani, 2021) relies on various factors, including GPT-2 perplexity, token sort ratio, and Levenshtein distance ratio, to evaluate correction quality. IMPARA (Maeda et al., 2022) fine-tunes BERT using only parallel data to quantify the impact of corrections. In terms of quality estimation, Chollampatt and Ng (2018a) introduced the first neural approach that does not rely on handcrafted features, while Liu et al. (2021) considered interactions between hypotheses using inference graphs.

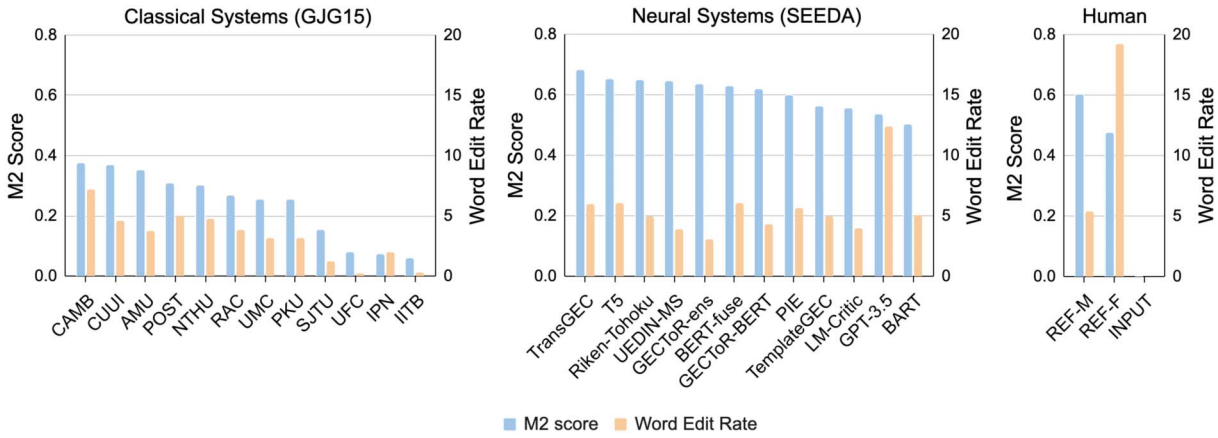


Figure 1: M^2 Score ($F_{0.5}$) and word edit rate for classical systems in GJG15, neural systems in SEEDA, and human sentences. These neural systems generate more edits and better corrections compared to classical systems.

3 The SEEDA Dataset

The SEEDA dataset consists of corrections annotated with human ratings along two different evaluation granularities (edit- and sentence-based), covering 12 state-of-the-art neural systems including LLMs, and two human corrections. The SEEDA dataset is denoted as SEEDA-E for edit-based evaluation and SEEDA-S for sentence-based evaluation. In this section, we describe the SEEDA dataset, how we generated the corrections (§3.1), and how we collected the annotations (§3.2). We use the CoNLL-2014 test set (Ng et al., 2014) as our input data, consisting of test essays and their error annotations. The test essays are written by non-native English-speaking students from the National University of Singapore and cover two genres: genetic testing and social media. Error annotations for the test essays are conducted by two native English speakers. The data comprises a total of 50 essays, consisting of 1,312 sentences and 30,144 tokens.

3.1 GEC Systems

To align with the current setting in GEC, we collect corrections using two mainstream neural-based approaches: *sequence-to-sequence* and *sequence tagging* (Bryant et al., 2023). To investigate how highly discriminating current metrics are, top-tier systems should be included among the target systems. This includes the LLMs that have received increased attention in recent years. Following these requirements, we carefully selected 11 systems, ensuring that the count is no less than the number of systems in GJG15. Among these, eight systems are sequence-to-sequence models that generate each token autoregressively: TemplateGEC (Li et al., 2023), TransGEC (Fang et al., 2023), T5 (Rothe et al., 2021), LM-Critic (Yasunaga et al., 2021), BART (Lewis et al., 2019), BERT-fuse (Kaneko et al., 2020), Riken Tohoku (Kiyono et al., 2019), and UEDIN-MS (Grundkiewicz et al., 2019). The remaining three systems are sequence tagging models that predict edit tags in parallel: GECToR-ens (Tarnavskyi et al., 2022), GECToR-BERT (Omelianchuk et al., 2020), and PIE (Awasthi et al., 2019). Following the recent LLMs trend, we consider GPT-3.5 (`text-davinci-003`) for two-shot learning (Coyne et al., 2023). We included INPUT (source from the CoNLL-2014 test set) since GEC evaluation requires consideration of uncorrected sentences. We also consider REF-M (minimal edit references by experts) and REF-F (fluency edit references by experts), which are introduced by Sakaguchi et al. (2016), to compare the system performance with human correction, bringing to the total to 15 sentence sets.

Figure 1 shows the M^2 Score ($F_{0.5}$)[2] and word edit rate for classical systems in GJG15, neural systems in SEEDA, and human sentences. Comparing these systems, neural systems in SEEDA

[2]In GEC, it is common to use $F_{0.5}$, where Precision is given twice the importance of Recall (Ng et al., 2014; Bryant et al., 2019). This is because, in the practical usage of GEC systems, not correcting is not as detrimental as making incorrect corrections. Additionally, in the context of language acquisition where minimizing incorrect feedback is desirable, this weighting is reasonable (Nagata and Nakatani, 2010).

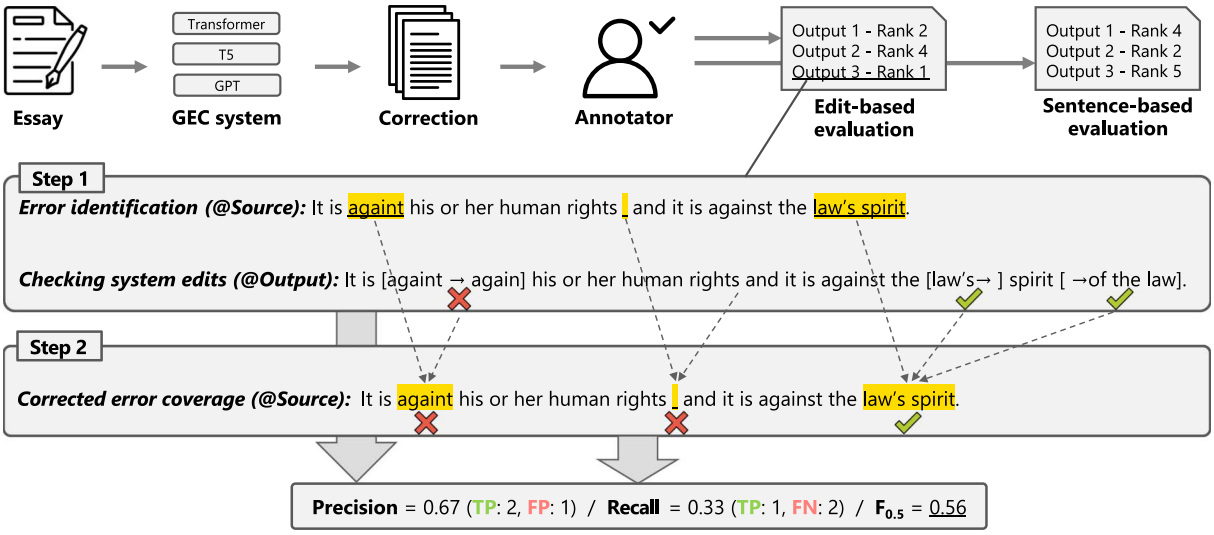


Figure 2: An overview of the annotation flow and an example of edit-based human evaluation. In Step 1, the annotator identifies errors in the source. Then, they categorize each edit in the output as either valid or not. In Step 2, the annotator determines whether each edit in the output effectively corrects the errors found in Step 1. TP, FP, and FN represent True Positive, False Positive, and False Negative, respectively.

show a higher number of edits and demonstrate better correction performance from the perspective of M^2. This performance comparison utilizes the most common GEC evaluation method, reproducing results reported in existing studies. On the other hand, this performance comparison contains intuitive contradictions, such as the lower performance of human-corrected sentences and LLMs. Therefore, we investigate and report how the modern system comparison deviates from human judgments (§4.2). Note that few-shot learning such as GPT-3.5 is known to be not grounding to target sentences as compared to finetuned models and may produce fluent but lengthy correction sentences that do not preserve the meaning of the source (Maynez et al., 2023).

3.2 Annotation Scheme

Edit-based Human Evaluation In the edit-based human evaluation, we evaluate only for edits in the system output. We perform a step-by-step sequence labeling using the doccano annotation tool (Nakayama et al., 2018). In the edit-based human evaluation, we decided to divide the process into two steps to avoid complicating the annotation process.

Figure 2 shows an overview of the annotation flow and an example of edit-based human evaluation. In Step 1, the detection of errors in the source and checking for edits in the output are performed. During the initial error detection, annotators refer to 25 error categories by Bryant et al. (2017) to identify error locations in the source, enabling them to label errors at the minimal unit level. In the subsequent Edit checking, annotators perform a binary decision to determine whether they would like to apply the edits in the output to improve the source or not. To reduce annotation costs, ERRANT is used for extracting edits. When there are conflicting edits (e.g., subject-verb agreement error), the one that aligns with the context is deemed effective, while the other is considered ineffective. Furthermore, for edits that depend on each other (e.g., [law's→] and [→of the law] in Figure 2), each is assigned an independent label, but they are deemed effective only if all dependent edits are present. In Step 2, the annotator performs a binary decision to determine whether each edit in the output effectively corrects the errors found in Step 1. Finally, we compute $F_{0.5}$ based on Precision and Recall[3] for each corrected sentence and subsequently rank the set of corrected sentences accordingly. The supplementary information about the annotation is provided in Appendix A.

Sentence-based Human Evaluation Following Grundkiewicz et al. (2015), sentence-based

[3]Note that Precision and Recall are computed at different levels of granularity.

#	Score	Range	System
1	0.273	1	AMU
2	0.182	2	CAMB
3	0.114	3–4	RAC
	0.105	3–5	CUUI
	0.080	4–5	POST
4	−0.001	6–7	PKU
	−0.022	6–8	UMC
	−0.041	7–10	UFC
	−0.055	8–11	IITB
	−0.062	8–11	INPUT
	−0.074	9–11	SJTU
5	−0.142	12	NTHU
6	−0.358	13	IPN

(a) Sentence-based evaluation in GJG15

#	Score	Range	System
1	0.992	1	REF-F
2	0.743	2	GPT-3.5
3	0.179	3–4	T5
	0.175	3–4	TransGEC
4	0.067	5–6	REF-M
	0.023	5–7	BERT-fuse
	−0.001	6–8	Riken-Tohoku
	−0.034	7–8	PIE
5	−0.163	9–12	LM-Critic
	−0.168	9–12	TemplateGEC
	−0.178	9–12	GECToR-BERT
	−0.179	9–12	UEDIN-MS
6	−0.234	13	GECToR-ens
7	−0.300	14	BART
	−0.231	13–14	BART
8	−0.992	15	INPUT

(b) Sentence-based evaluation in SEEDA

#	Score	Range	System
1	0.679	1	REF-F
2	0.583	2	GPT-3.5
3	0.173	3	TransGEC
4	0.097	4–6	T5
	0.078	4–7	REF-M
	0.067	4–7	Riken-Tohoku
	0.064	4–7	BERT-fuse
5	−0.076	8–11	UEDIN-MS
	−0.084	8–11	PIE
	−0.092	8–11	GECToR-BERT
	−0.097	8–11	LM-Critic
6	−0.154	12–12	GECToR-ens
7	−0.211	13–14	TemplateGEC
8	−0.797	15	INPUT

(c) Edit-based evaluation in SEEDA

Table 2: Human rankings for each evaluation granularity using TS. Systems based on GPT and T5 architectures (GPT-3.5, T5, TransGEC) consistently achieve higher rankings than REF-M, suggesting the potential for these systems to outperform human capabilities in providing corrections.

human evaluation is performed using the Appraise evaluation scheme (Federmann, 2010). Annotators read the context in the same way as edit-based human evaluation. And then, the corrected sentences are relatively ranked, allowing the same rank from the best to the worst. The judgment of whether a sentence is good or bad is left to the subjectivity of each annotator.

Annotator and Sampling Method Each annotation was performed by three native English speakers with extensive knowledge of the language. To observe differences by evaluation granularity, they are responsible for the same set of edit-based and sentence-based annotations. Following Grundkiewicz et al. (2015), we sample 200 subsets from the 1312 correction sets against the CoNLL-2014 test set using a parameterized distribution that favors more diverse outputs. To measure inter- and intra-annotator agreements, we duplicated at least 12.5% of the subset. One subset may contain up to five sentences, and the annotator creates a ranking from those sentences.

4 Dataset Analysis

In this section, we analyze SEEDA with a focus on evaluation granularity. First, we present the dataset statistics (§4.1). Second, we produce human rankings for the system using rating algorithms to conduct system-level meta-evaluation (§4.2). Third, we quantitatively analyze to discern any disparities in human evaluations across different evaluation granularities (§4.3).

Annotator	Raw data	Expanded
1	1,777 (592 / 507)	10,893 (6,349 / 5,919)
2	1,770 (522 / 240)	11,663 (7,053 / 5,445)
3	1,800 (343 / 44)	10,988 (5,572 / 4,433)
Total	5,347 (1,457 / 791)	33,544 (18,974 / 15,797)

Table 3: Dataset statistics for pairwise judgments by annotators. The numbers within the parentheses represent the number of ties, with the left being edit-based and the right being sentence-based.

4.1 Dataset Statistics

Table 3 presents the statistics for pairwise judgments by annotators. Each annotator has created 200 rankings for each subset, resulting in a total of 600 rankings. We take all combinations of all two sentences (A, B) for ranking, make a pairwise judgment (A>B, A=B, A<B), and count their numbers. To investigate the frequency of duplicate corrections, the raw data was expanded by treating systems that produced the same output independently. As a result, the number of pairwise evaluations increased significantly. This finding, similar to classical systems in Grundkiewicz et al. (2015), suggests that even high-performing neural systems that make many edits often generate

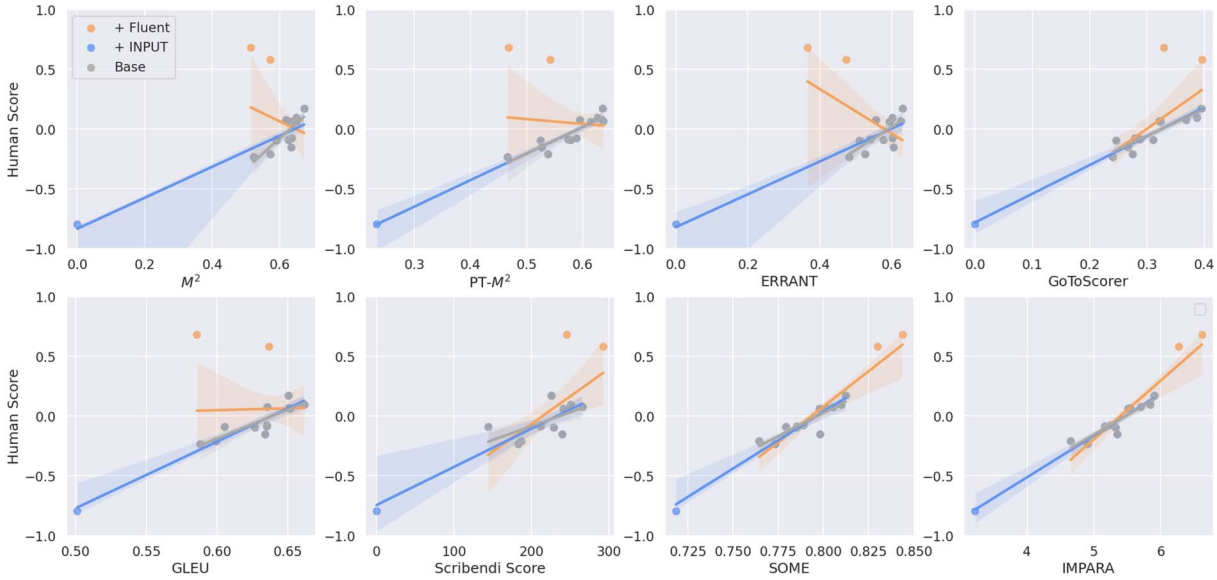


Figure 3: Scatter plots of the human score and the metric score. ‘‘Base’’ indicates the 12 systems excluding uncorrected sentences (INPUT) and fluent sentences (REF-F, GPT-3.5). Each line represents a regression line, and the shaded area indicates the size of the confidence interval for the estimated regression, obtained using bootstrap. Comparing the orange and blue regression lines to the gray regression line allows us to observe the degree of influence of each outlier on the distribution trend. For example, the leftward tilt of the orange regression lines for M^2, PT-M^2, ERRANT, and GLEU indicates a negative impact from fluent sentences as outliers.

Agreement	κ (SEEDA /GJG15)	Degree
Inter- (Edit)	0.28 / –	Fair
Inter- (Sentence)	0.41 / 0.29	Moderate
Intra- (Edit)	0.61 / –	Substantial
Intra- (Sentence)	0.71 / 0.46	Substantial

Table 4: Cohen’s κ measures the average inter- and intra-annotator agreements on pairwise judgments. The numbers in parentheses represent the κ for GJG15.

duplicated corrections. Moving forward, experiments will be conducted using raw data of pairwise judgments. Table 4 shows average inter- and intra-annotator agreements. Cohen’s kappa coefficient (κ) (Cohen, 1960) is used to measure the agreement. In comparison to the results in Grundkiewicz et al. (2015), the high inter- and intra-annotator agreement indicates that the annotators were able to provide more consistent evaluations.

4.2 Human Rankings

Following Grundkiewicz et al. (2015), we employed two rating algorithms, TrueSkill (TS) from Sakaguchi et al. (2014) and Expected Wins (EW) from Bojar et al. (2013), to create human rankings based on pairwise judgments. Table 2 shows the human rankings generated using TS for both edit-based and sentence-based evaluations. In contrast to classical systems in GJG15, all the neural systems receive ranks surpassing INPUT. This indicates a tendency of these systems to improve uncorrected sentences through correction. Systems based on GPT and T5 architectures (e.g., GPT-3.5, T5, TransGEC) achieve higher rankings than REF-M. This suggests the potential of these systems to offer corrections that might even surpass human capabilities.

4.3 Difference in Human Evaluation by Granularity

We perform a quantitative analysis of the variations in human evaluation based on granularity. To measure sentence-level agreement, we calculate the average intra-annotator κ between edit-based and sentence-based evaluations. The result, a modest 0.36, indicates low agreement. On the other hand, the system-level κ using pairwise judgments from the human rankings stands at a much higher 0.83, revealing negligible disparity. This indicates

a pronounced difference in sentence-level evaluation, but a relatively minor one in system-level evaluation. This suggests that biases are more prominent at the sentence-level meta-evaluation.

5 Baseline Metrics

We target 11 GEC metrics for meta-evaluation, including EBMs (§5.1) and SBMs (§5.2).

5.1 Edit-based Metrics

M^2 (Dahlmeier and Ng, 2012). It compares the edits in the corrected sentence with those in the reference. It dynamically searches for edits to optimize alignment with the reference edits using Levenshtein alignment (Levenshtein, 1966).

SentM^2. It is a variant of M^2 that calculates $F_{0.5}$ score at the sentence level.

PT-M^2 (Gong et al., 2022). It is a hybrid metric that combines M^2 and BERTScore (Zhang et al., 2019). It can measure the semantic similarity between pairs of sentences, not just comparing edits.

ERRANT (Bryant et al., 2017). It is similar to M^2 but differs in that it uses linguistically enhanced Damerau-Levenshtein alignment for extracting edits. It is characterized by its ability to calculate $F_{0.5}$ score for each error type.

SentERRANT. It is a variant of ERRANT that computes sentence-level $F_{0.5}$ score.

PT-ERRANT. It is a variant of PT-M^2 where the base metric has been changed from M^2 to ERRANT.

GoToScorer (Gotou et al., 2020). It calculates $F_{0.5}$ score while considering the difficulty of correction. The difficulty is calculated based on the number of systems that were able to correct the error.

5.2 Sentence-based Metrics

GLEU (Napoles et al., 2015). It is based on the commonly used BLEU (Papineni et al., 2002) in machine translation. It rewards N-grams in the output that match the reference but are not in the source while penalizing N-grams in the source that do not match the reference. For better evaluations, we use GLEU without tuning (Napoles et al., 2016a).

Scribendi Score (Islam and Magnani, 2021). It evaluates by combining the perplexity calculated by GPT-2 (Radford et al., 2019), token sort ratio, and Levenshtein distance ratio.

SOME (Yoshimura et al., 2020). It optimizes human evaluations by fine-tuning BERT separately for each of the following criteria: grammaticality, fluency, and meaning preservation.

IMPARA (Maeda et al., 2022). It combines a quality estimation model fine-tuned with parallel data using BERT and a similarity model to consider the impact of edits.

6 Revisiting Meta-evaluation for GEC

We investigate how correlations are affected by resolving granularity inconsistencies and are changed from classical systems to modern neural systems through system-level (§6.1) and sentence-level (§6.2) meta-evaluations. Figure 3 shows the scatter plots of the human evaluation and the metric scores, indicating that uncorrected sentences (INPUT) and fluently corrected sentences (REF-F, GPT-3.5) stand out as outliers and influence the correlation. Therefore, we consider 12 systems, deliberately excluding uncorrected sentences (INPUT) and sentences with fluently corrected sentences (REF-F, GPT-3.5). We calculate metric scores on the subset targeted in human evaluations.

6.1 System-level Meta-evaluation

Setup For our system-level meta-evaluation, we report correlation using system scores obtained from human rankings. Metrics such as M^2, PT-M^2, ERRANT, GoToScorer, and GLEU can calculate system scores, while other metrics use the average of sentence-level scores as the system score. We use Pearson correlation (r) and Spearman rank correlation (ρ) to measure the closeness between the metric and human evaluation.

Result According to the system-level meta-evaluation results in Table 5, it is evident that aligning the granularity between the metrics and human evaluation improves the correlation for EBMs in SEEDA-E, while the correlation for SBMs in SEEDA-S tends to decrease. One reason for the inconsistent results even when the granularity is aligned is that system-level human evaluations exhibit relatively small variations across different evaluation granularities.

Metric	System-level						Sentence-level					
	GJG15		SEEDA-S		SEEDA-E		GJG15		SEEDA-S		SEEDA-E	
	r	ρ	r	ρ	r	ρ	Acc	τ	Acc	τ	Acc	τ
M^2	0.721	0.706	0.658	0.487	0.791	0.764	0.506	0.350	0.512	0.200	0.582	**0.328**
SentM^2	0.852	0.762	0.802	0.692	0.887	0.846	0.506	0.350	0.512	0.200	0.582	**0.328**
PT-M^2	0.912	0.853	0.845	0.769	0.896	0.909	**0.512**	0.354	**0.527**	**0.204**	**0.587**	0.293
ERRANT	0.738	0.699	0.557	0.406	0.697	0.671	0.504	**0.356**	0.498	0.189	0.573	0.310
SentERRANT	0.850	0.741	0.758	0.643	0.860	0.825	0.504	**0.356**	0.498	0.189	0.573	0.310
PT-ERRANT	**0.917**	**0.886**	0.818	0.720	0.888	0.888	0.493	0.343	0.497	0.158	0.553	0.246
GoToScorer	0.691	0.685	**0.929**	**0.881**	**0.901**	**0.937**	0.336	0.237	0.477	−0.046	0.521	0.042
GLEU	0.653	0.510	0.847	**0.886**	**0.911**	0.897	0.684	0.378	0.673	0.351	0.695	0.404
Scribendi Score	0.890	0.923	0.631	0.641	0.830	0.848	0.498	0.009	0.354	−0.238	0.377	−0.196
SOME	**0.975**	**0.979**	0.892	0.867	0.901	**0.951**	**0.776**	**0.555**	**0.768**	**0.555**	**0.747**	**0.512**
IMPARA	0.961	0.965	**0.911**	0.874	0.889	0.944	0.744	0.491	0.761	0.540	0.742	0.502

Table 5: System-level and sentence-level meta-evaluation results excluding outliers. We use Pearson (r) and Spearman (ρ) for system-level and Accuracy (Acc) and Kendall (τ) for sentence-level meta-evaluations. The sentence-based human evaluation dataset is denoted SEEDA-S and the edit-based one is denoted SEEDA-E. The score in bold represents the metrics with the highest correlation at each granularity. There is a trend of improving correlation by aligning the metrics at the sentence level (SEEDA-S vs SEEDA-E) and a trend of decreasing correlation by changing the target systems from classical systems to neural systems (GJG15 vs SEEDA-S).

We discovered that as we move from classical systems to neural systems, correlations for all metrics—except GoToScorer and GLEU—decrease through a comparison between GJG15 and SEEDA-S. This result suggests that the majority of current metrics cannot adequately evaluate the more extensively edited and fluent corrections produced by neural systems, in contrast to those generated by classical systems. In the meta-evaluation results of GJG15, comparing it with existing studies (Grundkiewicz et al., 2015; Choshen and Abend, 2018a) is unfeasible, as the exclusion of INPUT has been implemented to alleviate scoring bias between EBMs and sentence-based human evaluation.

6.2 Sentence-level Meta-evaluation

Setup In sentence-level meta-evaluation, we use pairwise judgments in Table 3 to calculate correlations. We use Kendall's rank correlation (τ) and Accuracy (Acc) to measure the performance of the metrics. Kendall (τ) can measure performance in the common use case of comparing corrected sentences to each other.

Result In contrast to the system-level results, sentence-level meta-evaluations showed more significant improvements in correlations when the granularity was aligned. The substantial variation in sentence-level human evaluations based on granularity likely contributed to more consistent results. In other words, it became evident that correlations in sentence-level meta-evaluation are underestimated when granularity is not aligned.

When we compared GJG15 and SEEDA-S, we observed a decrease in correlations for most metrics, especially in EBMs, similar to the system-level results. Consistently high correlations were found for SOME and IMPARA, indicating the effectiveness of fine-tuned BERT.

7 Further Analysis

As further analysis, we investigate the influence of outliers (§7.1) and variations in the system set (§7.2) on the correlation of the metric. We test the hypothesis on which this study focuses, that there may be a range of correlations in flexible settings in GEC. Based on the best practices obtained in §6, granularity will be aligned in subsequent meta-evaluations.

7.1 Influence of Outliers

Table 6 shows the results when the uncorrected sentences (INPUT) and/or fluently corrected sentences (REF-F, GPT-3.5) are added to the base meta-evaluation excluding outliers (§6).

Metric	System-level						Sentence-level					
	+INPUT		+REF-F, GPT-3.5		All systems		+INPUT		+REF-F, GPT-3.5		All systems	
	r	ρ	r	ρ	r	ρ	Acc	τ	Acc	τ	Acc	τ
M^2	0.928	0.814	–0.239	0.161	0.566	0.318	0.605	0.361	0.527	0.216	0.558	0.264
M^2 (+Min)	0.929	0.884	–0.172	0.264	0.587	0.403	0.673	**0.461**	0.594	0.304	0.630	0.363
M^2 (+Min, Flu)	0.930	0.880	–0.149	0.262	0.594	0.400	**0.674**	0.458	**0.595**	**0.305**	**0.631**	**0.364**
SentM^2	0.971	0.879	–0.062	0.358	0.542	0.479	0.605	0.361	0.527	0.216	0.558	0.264
PT-M^2	0.974	0.929	–0.083	0.442	0.509	0.546	0.608	0.332	0.542	0.200	0.571	0.250
ERRANT	0.925	0.742	–0.502	0.051	0.404	0.229	0.597	0.344	0.511	0.188	0.542	0.236
ERRANT (+Min)	0.922	0.753	–0.462	0.112	0.475	0.279	0.609	0.350	0.530	0.184	0.550	0.218
ERRANT (+Min, Flu)	0.920	0.725	–0.460	0.090	0.484	0.261	0.605	0.348	0.523	0.175	0.541	0.207
SentERRANT	0.965	0.863	–0.357	0.200	0.354	0.350	0.597	0.344	0.511	0.188	0.542	0.236
PT-ERRANT	0.972	0.912	–0.324	0.240	0.352	0.382	0.580	0.292	0.500	0.144	0.532	0.199
GoToScorer	**0.974**	**0.951**	**0.667**	**0.916**	**0.817**	**0.932**	0.468	–0.064	0.505	0.009	0.476	–0.048
GLEU	0.957	0.911	–0.039	0.475	0.453	0.574	0.698	0.400	0.611	0.227	0.639	0.285
GLEU (+Min)	0.868	**0.942**	0.236	0.704	0.593	0.760	0.758	0.519	0.662	0.327	0.685	0.372
GLEU (+Min, Flu)	0.857	0.935	0.275	0.700	0.610	0.756	0.756	0.513	0.727	0.463	0.684	0.370
Scribendi Score	0.902	0.718	0.611	0.717	0.755	0.770	0.316	–0.323	0.345	–0.264	0.315	–0.328
SOME	0.965	0.896	0.931	0.916	**0.947**	0.932	**0.792**	**0.601**	**0.760**	**0.531**	**0.766**	**0.537**
IMPARA	**0.975**	0.901	**0.932**	**0.921**	0.934	**0.936**	0.785	0.587	0.742	0.496	0.745	0.495

Table 6: Meta-evaluation results when an outlier is included. Green indicates an increase in correlation compared to the meta-evaluation in Table 5, while red indicates a decrease. ‘‘+Min’’ in parentheses is when 11 minimal edit references are added, and ‘‘+Flu’’ is when three fluency edit references are added. ‘‘All systems’’ is the case where all outliers are considered. For most metrics, INPUT acts as an outlier that improves correlation, while REF-F and GPT-3.5 function as outliers that decrease correlation.

System-level Analysis The system-level results show that simply considering INPUT increases the correlations for most metrics to the point where comparisons are difficult. This suggests that INPUT serves as a strong outlier that skews the correlation positively and prevents accurate meta-evaluation. One of the reasons is that most EBMs assign the lowest score to INPUT, which also ranks the lowest in human evaluations. Therefore, in the meta-evaluation using neural models, it was demonstrated that a fair comparison cannot be made when considering the INPUT.

On the other hand, the addition of REF-F and GPT-3.5 shows a sharp drop in overall correlation. The results suggest that metrics other than SOME and IMPARA cannot properly assess fluently corrected sentences. Increasing references to commonly used metrics (M^2, ERRANT, GLEU) improves the correlation slightly, but still does not provide the same evaluation as humans. The same tendency as in the Maynez et al. (2023) study was observed that the overlap-based metric does not correctly evaluate LLMs for few-shot learning.

Sentence-level Analysis The results in the sentence-level meta-evaluation showed a similar trend as system-level results but with some differences. Adding INPUT improved correlations for most metrics, but both GoToScorer and Scribendi Score have decreased, which may be attributed to the inability to properly perform sentence-based evaluation. Furthermore, when adding REF-F and GPT-3.5, not only did many metrics show a decrease in correlation, but SOME and IMPARA also exhibited a slight reduction in correlation.

The improved correlations in M^2 (+Min) and GLEU (+Min), when REF-F and GPT-3.5 were added, indicate that the fluency correction may no longer be an outlier for commonly used metrics if the low coverage of reference-based evaluation is mitigated. To address the issue of reference coverage, an approach similar to Choshen and Abend (2018a), which involves splitting and combining edits for each reference, could potentially enhance the effective utilization of references. However, the result that fluency edit references were useful only for GLEU suggests that fluent edit references may be effective on an N-gram basis, but not on an edit extraction basis. As one of the reasons, we can consider the difficulties and complexities in edit extraction for fluent sentences in EBMs, as well as the inability to address the low coverage of three fluent references.

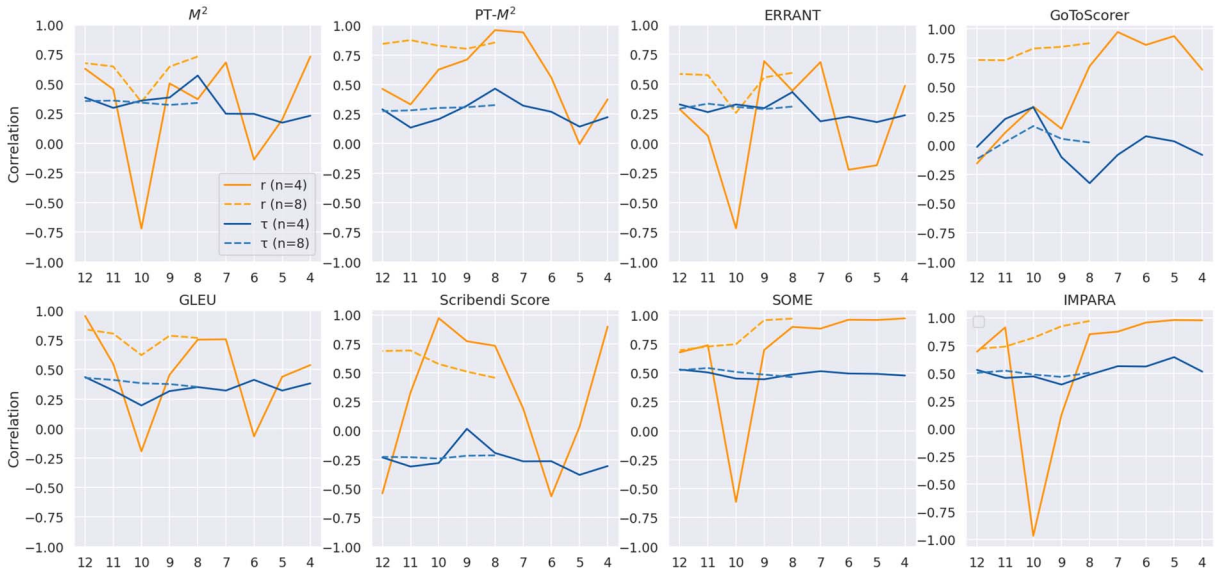


Figure 4: Variation of correlation when different systems are considered using window analysis. The x-axis represents the human ranking of the 12 systems excluding outliers. ‘‘n’’ denotes the number of systems considered, with solid lines representing four systems and dashed lines representing eight systems. For example, for $\mathrm{n} = 4$, a point with $x = 5$ corresponds to a human evaluation using systems ranked 2 to 5. The orange line represents Pearson (r) and the blue line represents Kendall (τ). The correlation of the main metrics (M^2, ERRANT, GLEU) shows significant variability, while pretraining-based metrics (SOME, IMPARA) exhibit relatively stable correlations.

7.2 Influence of Variations in the System Set

Next, we investigate the extent to which the correlation of the metrics varies with changes in a system set. To create a difficult setting for the metric, correlations are computed for a set of systems with close performance by sorting the systems in order of human ranking. Figure 4 shows the variation in correlations using window analysis. What is common for most metrics is that Pearson (r) tends to be highly variable from positive to negative for evaluation of four systems, but relatively stable for evaluation of eight systems. This suggests that most metrics do not have enough precision to identify performance differences in a set of high-performance neural systems. Therefore, there is still a need to develop better metrics that allow precise evaluation. Furthermore, M^2, ERRANT, and GLEU were often uncorrelated or negatively correlated, indicating that the commonly used metrics do not have high robustness. On the other hand, the BERT-based metrics were found to maintain relatively high correlations, with SOME in particular being the most robust. Kendall (τ) has a large number of samples for pairwise judgments, so there is no significant change.

8 Discussion

We provide a more practical guideline for meta-evaluation (§8.1) and evaluation (§8.2) methodologies in future GEC research by considering the experimental results so far.

8.1 Towards Valid Meta-evaluation in GEC

We recommend that meta-evaluation be conducted at each evaluation granularity in GEC. Specifically, EBMs should use SEEDA-E, and SBMs should use SEEDA-S. The meta-evaluation using SEEDA should use the 12 systems as a baseline, excluding outliers, and add REF-F and GPT-3.5 if one wants to find out how well the metrics can evaluate fluent corrections. This allows meta-evaluation for the modern neural system without the bias of the granularity. Additionally, conducting experiments with various methodologies is crucial to validate the characteristics of metrics. Therefore, experiments using GMEG-Data for domain-specific meta-evaluation of SBMs and meta-evaluation by MAEGE, irrespective of granularity, should be considered if resources permit.

The further analysis in §7, which yielded results unavailable in §6, demonstrates that conducting meta-evaluation for only a single setting is inadequate in GEC. Therefore, it is necessary to measure correlations across multiple experimental settings, considering the presence of outliers and more realistic sets of systems with similar performance. Additionally, achieving meta-evaluation reliability in GEC using confidence intervals for correlations, like Deutsch et al.'s (2021) study, is considered important. Furthermore, annotation based on Multidimensional Quality Metrics (Lommel et al., 2014) can take into account error types and severity, potentially providing interesting insights when compared to results from WMT (Freitag et al., 2021, 2022).

8.2 Best Practices for GEC Evaluation

We recommend the use of both EBMs and SBMs in GEC. In light of the trend toward more fluent correcting systems such as the GPT model, the current combination of the CoNLL-2014 test set and M^2 will no longer be adequate for proper evaluation. Therefore, it is essential to use high correlation metrics, such as SOME or IMPARA, in addition to M^2, to enable the evaluation of LLMs and achieve a more human-like and robust evaluation. Alternatively, exhaustive fluency references should be prepared to improve M^2 correlations, or datasets such as JFLEG (Napoles et al., 2017) that can account for fluency, should be used. Furthermore, using LLMs, as reported in recent studies (Chiang and Lee, 2023; Liu and Fabbri, 2023; Kocmi and Federmann, 2023) as an effective evaluator for other generative tasks, may also prove beneficial in GEC. If resources allow, it would be good to conduct additional human evaluations. EBMs and SBMs each have different strengths. EBMs can calculate Precision, Recall, and F-score, allowing a detailed evaluation of the system performance. In terms of second language acquisition, the evaluation of each edit provides information about the error location, type, and amount, which can improve the quality of feedback and learning efficiency. Most SBMs, on the other hand, can evaluate without references, circumventing the problem of underestimating corrections that are limited by the coverage of references. Also, unlike EBMs, SBMs do not automatically give the lowest score to uncorrected sentences. This allows for a quantifiable measurement to determine whether a sentence has been improved or worsened as a result of correction.

9 Conclusion

To address issues in conventional meta-evaluation in English GEC, we construct a meta-evaluation dataset (SEEDA) consisting of corrections with human ratings along two different evaluation granularities, covering 12 state-of-the-art system corrections including LLMs, and two human corrections with different focuses. The dataset analysis reveals that the results of sentence-level human evaluation differ between granularities and that GEC systems based on GPT and T5 can correct as well as or better than humans. Also, through meta-evaluation using SEEDA, we demonstrate that EBMs may be underestimated in existing meta-evaluations and that matching the evaluation granularity of metrics with human evaluations tends to improve sentence-level correlations. By further analysis, we discovered the uncertainty of conclusions based on a single correlation and found that most metrics lacked the precision to distinguish differences among high-performance neural systems. Finally, we propose a methodology for meta-evaluation and evaluation in GEC. We hope that this paper contributes to further advancements in GEC.

Acknowledgments

We would like to express our gratitude to the action editor and anonymous reviewers for their constructive feedback. We also thank the annotators who contributed to building the dataset.

References

Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. 2017. Reference-based metrics can be replaced with reference-less metrics in evaluating grammatical error correction systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–348, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. 2019.

Parallel iterative edit models for local sequence transduction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4260–4270, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1435

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3302

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4406

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, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1074

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, pages 1–59. https://doi.org/10.1162/coli_a_00478

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.870

Shamil Chollampatt and Hwee Tou Ng. 2018a. Neural quality estimation of grammatical error correction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2528–2539, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1274

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

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, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1127

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, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2020

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46. https://doi.org/10.1177/001316446002000104

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the performance of GPT-3.5 and GPT-4 in grammatical error correction. abs/2303.14342.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62, Montréal, Canada. Association for Computational Linguistics.

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

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146. https://doi.org/10.1162/tacl_a_00417

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Tao Fang, Xuebo Liu, Derek F. Wong, Runzhe Zhan, Liang Ding, Lidia S. Chao, Dacheng Tao, and Min Zhang. 2023. TransGEC: Improving grammatical error correction with translationese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3614–3633, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.223

Christian Federmann. 2010. Appraise: An open-source toolkit for manual phrase-based evaluation of translations. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta. European Language Resources Association (ELRA).

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, Colorado. Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1060

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid), Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Peiyuan Gong, Xuebo Liu, Heyan Huang, and Min Zhang. 2022. Revisiting grammatical error correction evaluation and beyond. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6891–6902, Abu Dhabi, United Arab Emirates, Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.463

Takumi Gotou, Ryo Nagata, Masato Mita, and Kazuaki Hanawa. 2020. Taking the correction difficulty into account in grammatical error correction evaluation. In *Proceedings of the 28th International Conference on Computational*

Linguistics, pages 2085–2095, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.188

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. 2015. Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1052

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4427

Md Asadul Islam and Enrico Magnani. 2021. Is this the end of the gold standard? A straightforward reference-less grammatical error correction metric. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3009–3015, Online and Punta Cana, Dominican Republic, Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.239

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2014. The AMU system in the CoNLL-2014 shared task: Grammatical error correction by data-intensive and feature-rich statistical machine translation. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 25–33, Baltimore, Maryland. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-1703

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4248–4254, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.391

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1236–1242, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1119

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

VI Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461. https://doi.org/10.18653/v1/2020.acl-main.703

Yinghao Li, Xuebo Liu, Shuo Wang, Peiyuan Gong, Derek F. Wong, Yang Gao, Heyan Huang, and Min Zhang. 2023. TemplateGEC: Improving grammatical error correction with detection template. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6878–6892, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.380

Yixin Liu and Alexander R. Fabbri. 2023. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. abs/2311.09184.

Zhenghao Liu, Xiaoyuan Yi, Maosong Sun, Liner Yang, and Tat-Seng Chua. 2021. Neural quality estimation with multiple hypotheses for grammatical error correction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5441–5452, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.429

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. https://doi.org/10.5565/rev/tradumatica.77

Koki Maeda, Masahiro Kaneko, and Naoaki Okazaki. 2022. IMPARA: Impact-based metric for GEC using parallel data. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3578–3588, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.448

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. 2023. Benchmarking large language model capabilities for conditional generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9194–9213, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.511

Ryo Nagata and Kazuhide Nakatani. 2010. Evaluating performance of grammatical error detection to maximize learning effect. In *Coling 2010: Posters*, pages 894–900, Beijing, China. Coling 2010 Organizing Committee.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Courtney Napoles, Maria Nădejde, and Joel Tetreault. 2019. Enabling robust grammatical error correction in new domains: Data sets, metrics, and analyses. *Transactions of the Association for Computational Linguistics*, 7:551–566. https://doi.org/10.1162/tacl_a_00282

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China. Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-2097

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

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016b. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2115, Austin, Texas. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1228

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. https://doi.org/10.18653/v1/E17-2037

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. https://doi.org/10.3115/v1/W14-1701

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.bea-1.16

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.89

Alla Rozovskaya, Kai-Wei Chang, Mark Sammons, Dan Roth, and Nizar Habash. 2014. The Illinois-Columbia system in the CoNLL-2014 shared task. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 34–42, Baltimore, Maryland. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-1704

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. https://doi.org/10.1162/tacl_a_00091

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3301

Maksym Tarnavskyi, Artem Chernodub, and Kostiantyn Omelianchuk. 2022. Ensembling and knowledge distilling of large sequence taggers for grammatical error correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3842–3852, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.266

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2021. LM-critic: Language models for unsupervised grammatical error correction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7763, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.611

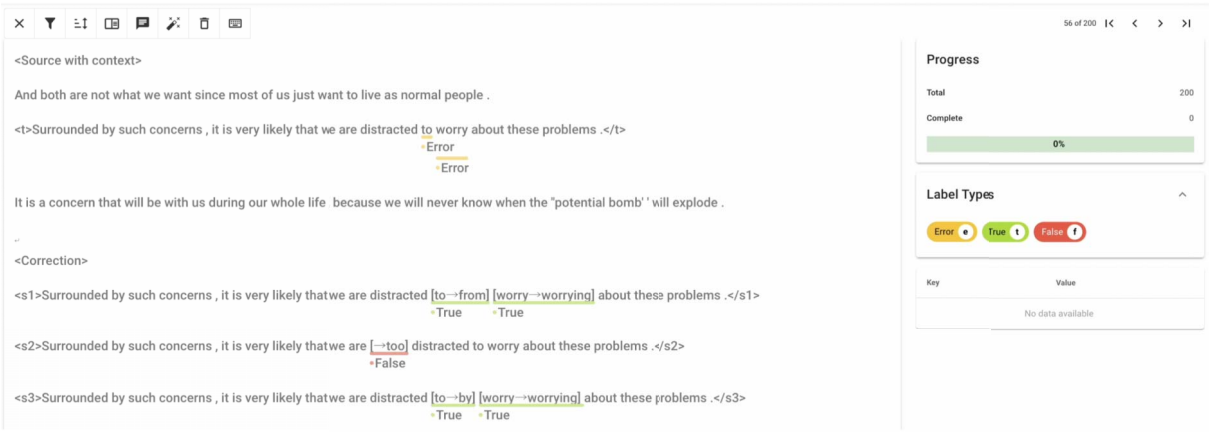
Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. SOME: Reference-less sub-metrics optimized for manual evaluations of grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*,

pages 6516–6522, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.573

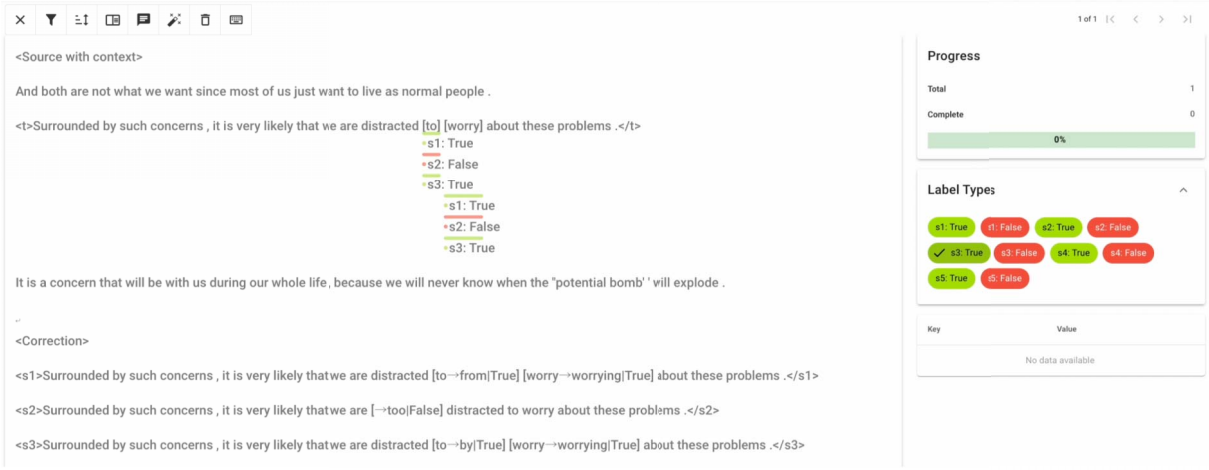
Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

A Supplement of Annotations

Figure 5 shows a screenshot of doccano used in the edit-based human evaluation. The source is enclosed in a <t> tag, and each corrected sentence is emphasized with a <s> tag along with the system number. In step 1, there are error labels for the source and True and False labels for each edit. In step 2, True and False labels with the system number are used to indicate whether the errors in the source were corrected. Due to the specifications of doccano, even if the same edit appears in multiple corrections, annotators need to label each occurrence separately. For information on Appraise in sentence-based human evaluation, you may refer to Grundkiewicz et al.'s (2015) work.



(a) Annotation step 1.



(b) Annotation step 2.

Figure 5: Screenshot of doccano used in the edit-based human evaluation.