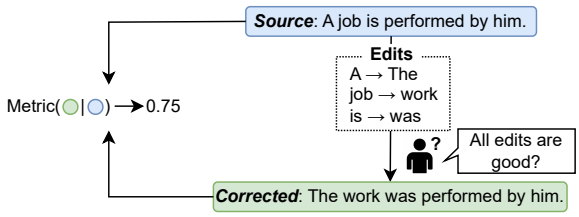
Improving Explainability of Sentence-level Metrics via Edit-level Attribution for Grammatical Error Correction

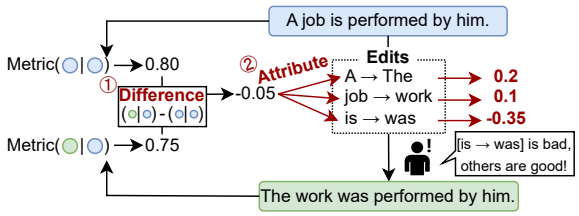
Takumi Goto, Justin Vasselli, Taro Watanabe
Nara Institute of Science and Technology
{goto.takumi.gv7, vasselli.justin_ray.vk4, taro}@is.naist.jp

Abstract

Various evaluation metrics have been proposed for Grammatical Error Correction (GEC), but many, particularly reference-free metrics, lack explainability. This lack of explainability hinders researchers from analyzing the strengths and weaknesses of GEC models and limits the ability to provide detailed feedback for users. To address this issue, we propose attributing sentence-level scores to individual edits, providing insight into how specific corrections contribute to the overall performance. For the attribution method, we use Shapley values, from cooperative game theory, to compute the contribution of each edit. Experiments with existing sentence-level metrics demonstrate high consistency across different edit granularities and show approximately 70% alignment with human evaluations. In addition, we analyze biases in the metrics based on the attribution results, revealing trends such as the tendency to ignore orthographic edits. Our implementation is available at GitHub.



(a) The existing metrics are low-explainability.



(b) Our proposed method improves explainability.

Figure 1: Overview of the proposed method with an example using three edits. Figure (a) shows the low-explainability of existing metrics that only estimate the sentence-level score, but Figure (b) shows that the edit-level attribution solves this issue by explaining which edit improves or worsens the sentence-level score.

1 Introduction

Grammatical error correction (GEC) is the task of automatically correcting grammatical or superficial errors in an input sentence. Automatic evaluation metrics play a key role in improving GEC performance, but their effectiveness depends on their level of explainability. For example, metrics that evaluate at the edit level are more explainable than sentence-level metrics, as they allow us to identify which specific edits are effective and which are not, even when a GEC system makes multiple edits. Such explainable metrics enable researchers to analyze the strengths and weaknesses of GEC models, providing valuable insights into how models can be improved. Furthermore, in education applications, explainable metrics can provide language learners with detailed feedback on their writing, supporting their learning more effectively.

In GEC, explainable reference-based metrics, such as ERRANT (Felice et al., 2016; Bryant et al., 2017) are limited because references cannot account for all valid corrections. Preparing test data with comprehensive references is often impractical, especially when targeting domains such as medical or academic writing that differ from existing datasets. To address this issue, reference-free metrics have been proposed to evaluate corrected sentences without relying on references (Choshen and Abend, 2018b; Yoshimura et al., 2020; Islam and Magnani, 2021; Maeda et al., 2022). Although these reference-free metrics achieve high correlation with human evaluations, many are designed to assign scores at the sentence level, limiting their explainability on individual edits. This lack of granularity makes it difficult to analyze how specific edits contribute to the overall sentence score. For example, as shown in Figure 1, a metric evaluates

a corrected sentence created by applying the three edits. As shown in Figure 1a, the sentence-level metric assigns an overall score of 0.75, but it does not indicate whether all edits are valid, or if both valid and invalid edits have been applied.

To improve the explainability of metrics with low or no explanation, we propose attributing sentence-level scores to individual edits as illustrated in Figure 1b. In our method, the total contribution of all edits is calculated as the difference between the scores of the input sentence and the corrected sentence. This difference is then attributed to the individual edits. In Figure 1b, a difference of -0.05 is distributed among three edits with contributions of 0.2, 0.1, and -0.35. The attribution results are interpreted using the sign and magnitude of these scores: the sign indicates whether an edit is valid or not, while the magnitude represents the degree of its influence on the final sentence-level score. We employ Shapley values (Shapley et al., 1953) from cooperative game theory to fairly distribute the total score among the edits. By considering all combination of edits, Shapley values allow us to precisely attribute each edit's contribution to the overall sentence score, offering insights into their individual impact. Unlike existing attribution methods which typically calculate contributions at the token level (Lundberg and Lee, 2017; Sundararajan et al., 2017), our novel approach computes contributions for changes in a sentence.

In the experiments, we apply our method to two popular reference-free metrics, SOME (Yoshimura et al., 2020) and IMPARA (Maeda et al., 2022), as well as a fluency metric based on GPT-2 (Radford et al., 2019) perplexity. The results show that the proposed attribution method assigns consistent scores across different granularities of edits and that edits with larger absolute attribution scores align more closely with human evaluations. We also introduce Shapley sampling values (Strumbelj and Kononenko, 2010) to mitigate the time-complexity issues of exact Shapley values. Additionally, we demonstrate that the proposed method can explain metric decisions at both the sentence and corpus levels, categorized by error types. These analyses reveal the types of edits that metrics give more weight to, as well as provide insights into the strengths and weaknesses of GEC systems.

2 Background

Edits in GEC. The GEC task aims to correct grammatical errors in a source sentence S and output a corrected sentence H. The differences between S and H are often represented as N edits $\boldsymbol{e} = \{e_i\}_{i=1}^{N}$ to enable evaluation (Dahlmeier and Ng, 2012; Bryant et al., 2017; Gong et al., 2022; Ye et al., 2023), ensembling (Tarnavskyi et al., 2022), and post-processing (Sorokin, 2022) at the edit level. These edits can be automatically extracted using edit extraction methods (Felice et al., 2016; Bryant et al., 2017; Belkebir and Habash, 2021; Korre et al., 2021; Uz and Eryiğit, 2023). Each edit typically includes a word-level span in S and its corresponding correction, although it may also include an error type (Bryant et al., 2017). The error type categorizes each edit, indicating the part-of-speech or grammatical aspect it relates to, which helps analyze the strengths and weaknesses of GEC systems.

Sentence-level Metrics. A sentence-level metric M computes the score of the corrected sentence given the source sentence, denoted as $M(H|S) \in \mathbb{R}$. The source sentence is used to assess meaning preservation, as GEC requires correcting errors while maintaining the original meaning of the source sentence. This formulation has been adopted by several reference-free metrics (Yoshimura et al., 2020; Islam and Magnani, 2021; Maeda et al., 2022; Kobayashi et al., 2024a). Sentence-level metrics aim to rank GEC systems in alignment with humans judgments, as evidenced by the fact that the meta-evaluation is performed using the correlation between metric-generated rankings or scores and those of humans. However, these metrics are limited to sentence-level scoring and cannot explain how individual edits contribute to the final score.

Edit-level Weighting Some metrics already employ edit-level weighting. GoToScorer (Gotou et al., 2020) weights edits using the correction success rate of a pre-defined GEC system set, while PT-ERRANT (Gong et al., 2022) weights based on the difference of BERTScore (Zhang et al., 2019) when applying and not applying an edit to the incorrect sentence. CLEME (Ye et al., 2023) weights edits according to their span length, and CLEME2.0 (Ye et al., 2024) uses the same weighting strategy as PT-ERRANT. The goal of GoToScorer is to promote error corrections that other systems cannot

correct, while the goal of PT-ERRANT, CLEME, and CLEME2.0 is to improve agreement with human evaluation results. MAEGE (Choshen and Abend, 2018a) is a preexisting meta-evaluation method which involves quantifying the contribution of edits to a score from a reference-based metric. Unlike MAEGE, our approach is grounded in the robust theory of Shapley values, and works on reference-free metrics.

3 Method

Our attribution method assumes that the overall contribution of edits is the difference in scores before and after correction. We distribute the difference $\Delta M(H|S) = M(H|S) - M(S|S)$ across each edit $\boldsymbol{e} = \{e_i\}_{i=1}^N$, where $M(S|S)$ is the score of the source sentence treated as its own corrected sentence.

The goal of our attribution method is to compute the contribution for each edit denoted as $\{\phi_i(M) \in \mathbb{R}\}_{i=1}^N$, so that the following equation is satisfied:

$$\Delta M(H|S) = \sum_{i=1}^{N} \phi_i(M). \tag{1}$$

We refer to $\phi_i(M)$ as *attribution scores*. A positive score ($\phi_i(M) > 0$) indicates an edit that improves the metric $M(\cdot)$, while a negative score ($\phi_i(M) < 0$) indicates an edit that worsens it. The absolute value $|\phi_i(M)|$ represents the degree of the edit's contribution. Unlike previous studies, e.g., GoToScorer and CLEME, the purpose of the attribution scores is to explain the internal decision of metrics.

Shapley. For the attribution method, we introduce Shapley values (Shapley et al., 1953) from cooperative game theory. In cooperative game theory, multiple players work together towards a common goal and share the total benefit based on their contributions. Shapley values distribute this benefit among players fairly, ensuring that those players who contributes more receive a larger share. For our purpose, we regard $\Delta M(H|S)$ as the total benefit, edits $\boldsymbol{e}$ as the players, and $\phi_i(M)$ as the Shapley values. The Shapley value $\phi_i(M)$ for a given metric $M(\cdot)$ is calculated as follows:

$$\phi_i(M) = \sum_{\boldsymbol{e}' \subseteq \boldsymbol{e} \setminus \{e_i\}} \frac{|\boldsymbol{e}'|!(N - |\boldsymbol{e}'| - 1)!}{N!} \left(\Delta M(S_{\boldsymbol{e}' \cup \{e_i\}}|S) - \Delta M(S_{\boldsymbol{e}'}|S)\right), \tag{2}$$

where $S_{\boldsymbol{e}}$ denotes the source sentence after applying the edit set $\boldsymbol{e}$. Equation 2 calculates the weighted sum of the differences in evaluation scores when including and excluding the edit e_i. For example, using Figure 1 with $\boldsymbol{e} = \{e_1, e_2, e_3\} = \{[\text{A} \to \text{The}], [\text{job} \to \text{work}], [\text{is} \to \text{was}]\}$, one of the terms in the calculation for $\phi_1(M)$ with $\boldsymbol{e}' = \{e_2\}$ is

$$\begin{aligned} &\frac{1}{6}\left(\Delta M(S_{\{e_1,e_2\}}|S) - \Delta M(S_{\{e_2\}}|S)\right) \\ &= \frac{1}{6}(\Delta M(\textbf{The } \underline{\text{work}} \ \underline{\text{is}} \text{ performed by him.}|S) \\ &- \Delta M(\textbf{A } \underline{\text{work}} \ \underline{\text{is}} \text{ performed by him.}|S)). \end{aligned} \tag{3}$$

Here, bold words indicate the edit being attributed, and underlined words show other edits. The terms for $\boldsymbol{e}' = \{\phi\}$, $\{e_3\}$, and $\{e_2, e_3\}$ are computed in a similar way. Shapley values consider various combinations of edits, ensuring accurately attribution of the i-th edit's contribution. By design, Shapley values naturally satisfy Equation 1 due to their *effectiveness* (Shapley et al., 1953). However, the computational complexity is $\mathcal{O}(2^N)$.

Shapley Sampling Values. To improve computational efficiency, we introduce Shapley sampling values (Strumbelj and Kononenko, 2010), an approximation of Shapley values. Equation 2 can be rewritten as:

$$\phi_i(M) = \frac{1}{N!} \sum_{\boldsymbol{o} \in \pi(\boldsymbol{e})} \left(\Delta M(S, S_{\mathrm{Pre}^i(\boldsymbol{o}) \cup \{e_i\}})) - \Delta M(S, S_{\mathrm{Pre}^i(\boldsymbol{o})})\right) \tag{4}$$

where $\pi(\boldsymbol{e})$ is the set of all possible orders of edits, and $\mathrm{Pre}^i(\boldsymbol{o})$ is the set of edits preceding e_i in permutation $\boldsymbol{o}$. In the example from Equation 3, $\mathrm{Pre}^1(\boldsymbol{o}) = \{\phi\}$ when $\boldsymbol{o} = [e_1, e_2, e_3]$, and $\mathrm{Pre}^1(\boldsymbol{o}) = \{e_2, e_3\} = \{[\text{job} \to \text{work}], [\text{is} \to \text{was}]\}$ when $\boldsymbol{o} = [e_3, e_2, e_1]$. To approximate Shapley values, we uniformly sample T permulations without replacement from $\pi(\boldsymbol{e})$, denoted as $\tilde{\pi}(\boldsymbol{e}) = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$. Shapley sampling values are then calculated using $\tilde{\pi}(\boldsymbol{e})$ instead of $\pi(\boldsymbol{e})$ in Equation 4. This approximation reduces the computational cost from $\mathcal{O}(2^N)$ to $\mathcal{O}(TN)$.

Normalized Shapley Values The calculated attribution scores are not directly comparable across different sentence-level scores. For instance, an attribution score of 0.2 has a different relative impact when distributing a sentence-level score of 1.0 versus 0.4. To enable meaningful comparison, we apply L1 normalization to the attribution scores:

$$\phi_i^{\text{norm}}(M) = \frac{\phi_i(M)}{\sum_{i=1}^{N} |\phi_i(M)|}. \quad (5)$$

This normalization, applied as a post-processing step, adjusts only the magnitude of the scores while preserving their original signs. Since the normalized scores represent the ratio of each edit's contribution, they are assumed to be comparable even when the sentence-level scores differ.

4 Evaluation of Attribution

We evaluate the proposed attribution method from two perspectives: faithfulness and explainability (Wang et al., 2024). Faithfulness measures how well the attribution results reflect the model's internal decision, while explainability assesses the extent to which the results are understandable to humans. To demonstrate the effectiveness of the proposed method across various domains, we conduct experiments using diverse datasets, GEC systems, and metrics.

4.1 Experimental Settings

4.1.1 Datasets

We use CoNLL-2014 test set (Ng et al., 2014) and the JFLEG validation set (Heilman et al., 2014; Napoles et al., 2017). CoNLL-2014 is a benchmark for minimal edits, focusing on correcting errors while preserving the original structure of the input as much as possible. In contrast, JFLEG is a benchmark for fluency edits, allowing more extensive rewrites to produce fluent and natural sentences.

4.1.2 GEC Systems

We evaluate our attribution method on various GEC systems, including two tagging-based models (the official RoBERTa-based GECToR (Omelianchuk et al., 2020) and GECToR-2024 (Omelianchuk et al., 2024)), two encoder-decoder models (BART (Lewis et al., 2020) and T5 (Rothe et al., 2021)), and a causal language model (GPT-4o mini). This allows us to assess the explainability of attributions scores across different GEC architectures. For GPT-4o mini, we used a two-shot setting following Coyne et al. (2023), with examples

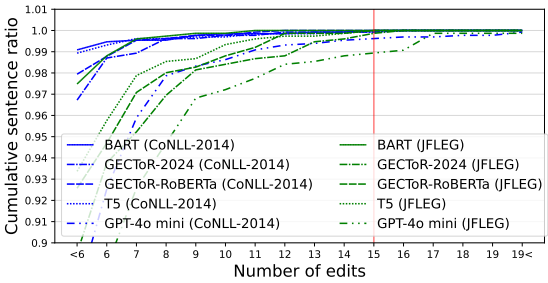


Figure 2: Cumulative sentences ratio regarding the number of edits. The red line indicates the position where the number of edits is 15.

randomly sampled once from the W&I+LOCNESS validation set (Yannakoudakis et al., 2018) and used for all input sentences. Note that we use only the corrected sentences containing 15 or fewer edits ($N \leq 15$) due to the computational complexity of Shapley values. According to Figure 2, which shows the cumulative sentence ratio regarding the number of edits, our experiments cover at least more than 98.9% of the sentences in all corrected sentences.

4.1.3 Reference-free Metrics

We use the following non-explainable metrics in the experiments. Other metrics such as reference-based metrics could also be used, but we do not use such already explainable metrics in this paper.

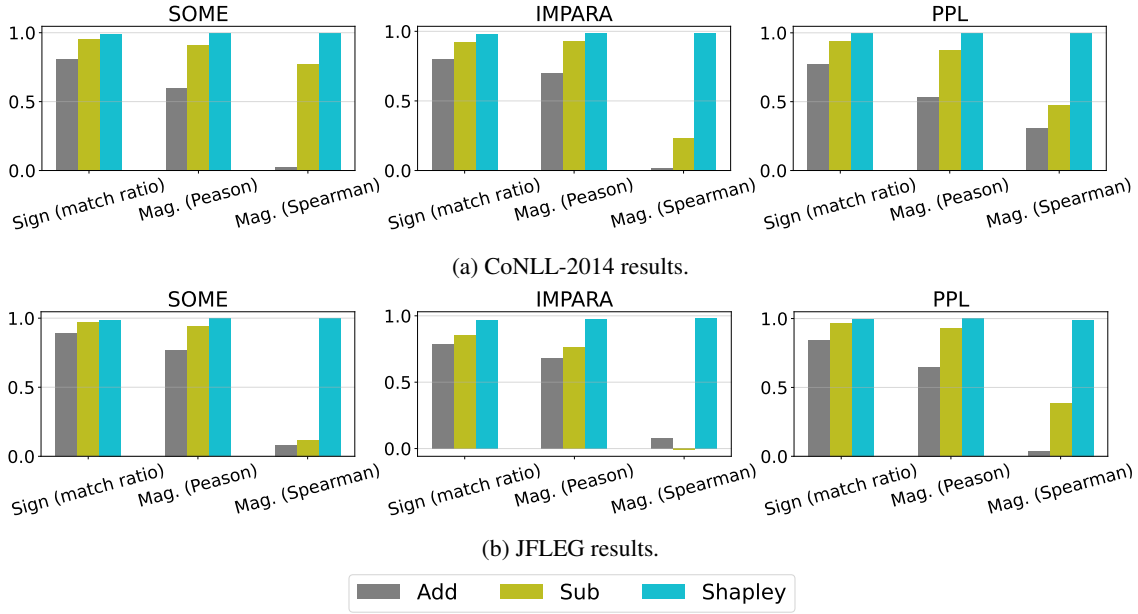
SOME (Yoshimura et al., 2020) uses a BERT-based regression model optimized directly on human evaluation results. We used the official pre-trained model weights[1] and used the default coefficients for the weighted average of grammaticality, fluency, and meaning preservation scores, from the official script[2].

IMPARA (Maeda et al., 2022) estimates evaluation scores through similarity estimation and quality estimation. We use BERT (`bert-base-cased`) as the similarity estimator and train our own model for the quality estimator, as the official pre-trained weights are not available. Our quality estimator was trained following the same settings described in Maeda et al. (2022), achieving a correlation with the human ranking comparable to their reported results.

GPT-2 Perplexity (PPL). Our proposed method can be applied to metrics that evaluate only the

[1] https://github.com/kokeman/SOME

[2] 0.55*grammaticality + 0.43 * fluency + 0.02 * meaning preservation.



(a) CoNLL-2014 results.

(b) JFLEG results.

Figure 3: The results of consistency-based evaluation. Each row shows the different datasets and each column shows different metrics. "Mag." means the magnitude. Colors show the attribution scores.

quality of the corrected sentence[3]. To test this, we use GPT-2 (Radford et al., 2019) perplexity, with negative perplexity scores to ensure that higher values correspond to better quality. Perplexity is one of the components employed in Scribendi score (Islam and Magnani, 2021).

4.2 Baseline Attribution Methods

To evaluate the effectiveness of Shapley values, we employ simpler variants, i.e., ADD and Sub, as baseline attribution methods.

Add. This method observes the change in the score when each edit is applied individually to the source sentence. An edit that increases the score is considered valid for the metric. This approach corresponds to using only $\boldsymbol{e}' = \{\phi\}$ in Equation 2, with the attribution scores normalized by $\frac{\Delta M(H|S)}{\sum_{i=1}^{N} \phi_i(M)}$ so that it satisfies Equation 1.

Sub. This method observes the change in the score when each edit is removed individually from the corrected sentence. An edit that decreases the score upon removal is considered valid for the metric. This approach corresponds to using only $\boldsymbol{e}' = \boldsymbol{e} \setminus \{e_i\}$ in Equation 2, with the attribution scores normalized by $\frac{\Delta M(H|S)}{\sum_{i=1}^{N} \phi_i(M)}$ so that it satisfies Equation 1.

[3]In this case, the sentence-level score is $\Delta M(S, H) = M(H) - M(S)$

4.3 Consistency Evaluation

To evaluate faithfulness, we test how well the attribution scores represent the judgments of the metrics through consistency evaluation. Specifically, we first calculate the attribution scores for individual edits and then group edits with the same sign, treating them as a single edit. Next, we calculate the attribution score for the grouped edits. We hypothesize that the attribution score for a grouped edit should equal the sum of the individual attribution scores of the edits comprising the group. If this condition holds, the attribution method consistently calculates the contributions of edits, making its results reliable for practical use. We use an agreement ration to measure the consistency of the signs and use Pearson and Spearman correlations to assess the consistency of the magnitudes.

For example, in Figure 1, we group two positivity-attributed edits, [*A* → *The*] and [*job* → *work*], into a single edit and compute attribution scores for the grouped edit and the remaining edit, [*is* → *was*]. Ideally, the attribution score for the grouped edit should be $0.2 + 0.1 = 0.3$, which can be verified by sign agreement and closeness to 0.3.

Figure 3 presents the results for each metrics. Our proposed Shapley method shows higher consistency than the baseline attribution methods across various domains and metrics. While the Sub metric also demonstrates high consistency, its Spearman's rank correlation occasionally drops for certain met-

rics, such as IMPARA. Low rank correlation can misrepresent the relative importance of edits, posing a serious issue for explainability. These results suggest that the attribution method is reliable across different edit granularities, such as edits extracted by ERRANT (Felice et al., 2016; Bryant et al., 2017) or chunks created by merging multiple edits (Ye et al., 2023). This flexibility enables a wide range of applications for the proposed method.

Metric	Error	Time	Shapley values dist.
SOME	0.014	3.86	0.019 ± 0.020
IMPARA	0.074	3.77	0.052 ± 0.071
PPL	19.610	0.82	34.549 ± 59.472

Table 1: The average error and average computation time (seconds) when using Shapley sampling values. It also shows the distribution of the absolute exact Shapley values (the average ± the standard deviation).

4.4 Human Evaluation

To evaluate explainability, we assess the agreement between attribution scores and edit-level human annotation in SEEDA (Kobayashi et al., 2024b), a meta-evaluation dataset based on CoNLL-2014. The annotation in SEEDA are represented as binary labels indicating whether an edit is valid or not. Ideally, a positively attributed edit should align with a valid edit in human evaluation, while a negativity attributed edit should align to an invalid one. We calculate accuracy at the corpus level by comparing the validity (valid/invalid) of annotation with the sign of attribution scores (positive/negative). SEEDA assigns one to five hypothesis sentences to each source sentence with each hypothesis annotated by three evaluators. We use the data corresponding to the first annotator, comprising 200 sources and 841 hypotheses [4].

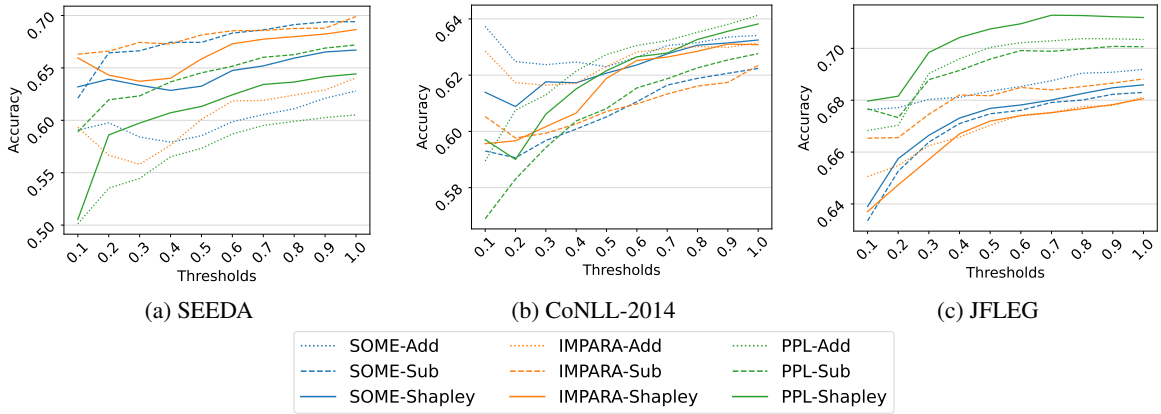
We also utilize a reference-based evaluation framework to approximately obtain human edit-level annotation. Evaluation with SEEDA are limited to CoNLL-2014 dataset and cannot be performed on data from other domains such as JFLEG, and newly annotating the edit-level validity is expensive. Sentence-level references are generally provided for many datasets, and approximately obtain edit-level human evaluation using the references. Specifically, we extract hypothesis edits given the source and hypothesis using ERRANT, in addition to reference edits given the source and reference. Then, we annotate a binary label to each hypothesis edit: valid if the edit is included in the reference edits, invalid otherwise. Here we use the official two references for CoNLL-2014 and four references for JFLEG. For each hypothesis, we select the one that has the highest accuracy with the attribution scores.

Although the above method approximately evaluates the sign of the attribution scores, it cannot evaluate the reliability of their magnitude. For the evaluation of magnitude, we follow standard attribution evaluation practices (Petsiuk, 2018; Fong and Vedaldi, 2017) by applying a threshold to the absolute values of the scores. To compute the agreement rate, we only consider edits whose normalized absolute attribution scores are below the specified threshold. The threshold starts at 0.1 and increases in steps of 0.1 until it reaches 1.0, where all edits are included. Ideally, the larger the threshold, the higher the accuracy, because more confidently attributed edits are used.

Figure 4 presents the results. Overall, the results show that including edits with larger absolute attribution scores improves the agreement with human evaluation, indicating that the magnitude of attribution scores is meaningful. Figure 4a at threshold=1.0 shows 60 % to 70% accuracy, which constantly agrees with the human evaluation considering that the random baseline is 50%. Figure 4b and Figure 4c also show a similar trend to Figure 4a, indicating that the use of direct human annotation can be replaced by the reference-based evaluation to investigate the agreement between attribution scores and human judgment.

When comparing attribution methods, Shapley rarely achieves the worst agreement. For instance, in JFLEG, SOME shows the order Add > Shapley > Sub, while IMPARA shows Sub > Shapley > Add. Either Add or Sub often results in the worst agreement, whereas Shapley demonstrates more stable performance across different metrics and domains. When comparing metrics, the rank order among metrics is reversed between directly annotated labels by humans and approximate labels by referential evaluation: IMPARA > SOME > PPL in Figure 4a, but PPL > SOME > IMPARA in Figure 4b and Figure 4c. There is a divergence in results between using direct and approximated labels. This suggests that using approximated labels might be inappropriate when discussing which metric yields the highest agreement with human evaluation.

[4] https://github.com/tmu-nlp/SEEDA/tree/main/data/EditEval_Step1/annotator1



(a) SEEDA (b) CoNLL-2014 (c) JFLEG

Figure 4: Human evaluation results. SEEDA directly uses evaluation results as human evaluation labels, while CoNLL-2014 and JFLEG use approximation labels extracted from references. The x-axis represents the threshold for attributed scores, and the y-axis indicates the agreement rate with the labels. A larger value on the x-axis indicates attribution scores with higher confidence.

Original (S)	-	Further more		by	these	evidence		u	will agree	
Correction (H)	-	Further more	,	with	this	evidence	,	you	will agree	.
Metrics (M)	$\Delta M(\cdot)$	Shapley values $\phi_i(M)$								
SOME	0.298	-	0.068	0.064	0.033	-	0.038	0.066	-	0.030
IMPARA	-0.027	-	0.068	0.029	0.124	-	0.145	-0.361	-	-0.033
PPL	1266.3	-	250.7	103.8	216.0	-	67.4	366.6	-	261.5
		Normalized Shapley values								
SOME		-	0.229	0.215	0.111	-	0.126	0.220	-	0.099
IMPARA		-	0.090	0.039	0.163	-	0.191	-0.475	-	-0.043
PPL		-	0.198	0.082	0.171	-	0.053	0.290	-	0.207

Table 2: An example of the proposed method's results using actual sentence.

4.5 Efficiency of Shapley Values

One limitation of Shapley values is their high computational cost. In our preliminary experiments using a single RTX 3090, we observed that the computation time reaches about 30 seconds when the number of edits in a corrected sentence exceeds 11. This observation shows that sentences with more than 11 edits are impractical to attribute within a reasonable time. As indicated by Figure 2, although only 3% of GEC outputs have more than 11 edits, those tasks involving a higher number of edits, e.g., text simplification, could face even greater challenges.

As discussed in Section 3, we address this issue by employing Shapley sampling values and evaluate their ability to approximate exact Shapley values by measuring the average absolute differences between them. In the experiments, we use a dataset combining all GEC model hypotheses on the JFLEG validation set. We set $T = 64$ and restrict examples to $10 \leq N \leq 15$ [5].

Table 1 reports average errors and computation times for each metric. With Shapley sampling values, the computation time per sentence can be reduced to as little as four second in average[6]. To assess the impact of errors, we also show the distribution of absolute exact Shapley values in Table 1. If the error exceeds the mean in this distribution, the likelihood of misunderstanding the contribution relationship between edits increases. While SOME and PPL show errors below the mean, IMPARA exhibits higher errors. IMPARA's higher error may be due to its smaller variance in evaluated values, making it less effective at quantifying impact with a limited number of calculations.

[5]When $T = 64$ and $10 \leq N$, the computation cost of Shapley sampling values is consistently lower than that of exact Shapley values, as $2^x > 64x$ holds for $x > 9.20\ldots$.

[6]Refers to Appendix A for more detailed results.

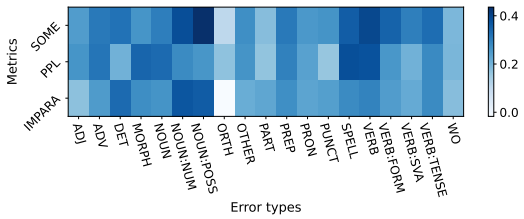


Figure 5: The heatmap indicating the average of normalized Shapley values per error type. The deeper color indicates higher values.

5 Applications of Attribution Scores

We demonstrate practical applications of attribution scores for users. All results in this section are based on Shapley values for the attribution method.

5.1 Case Study

Attribution scores can be used to identify which edits improve or worsen the sentence-level score. Table 2 provides an example, showing attribution scores and their normalized version. The original sentence and its corrections are chunked according to edit spans, omitting scores for non-edited chunks which are all zeros. One observation is that the sentence-level score of IMPARA declines primarily due to the edit [$u \rightarrow you$], which is inconsistent with human intuition. In contrast, SOME and PPL prefer this edit. This observation of IMPARA suggests a problem with IMPARA's scoring, does not imply a problem with our attribution method, and rather it reveals weaknesses in metrics through case studies.

Normalized Shapley values enable comparison of attribution scores across metrics. For example, while SOME and IMPARA assign the same Shapley value to the edit [$\phi \rightarrow ,$], their normalized scores reveal different impacts. This feature is particularly useful for comparing metrics with different value ranges, such as SOME and PPL.

Beyond case studies, we also investigate metric bias at the corpus level. To investigate these biases, we calculate the average normalized Shapley values for each error type (Bryant et al., 2017). We merge the corrected sentences from five GEC systems for the JFLEG validation set to mitigate biases specific to individual GEC models. Figure 5 shows the results for error types with a frequency greater than 30 and indicate that different metrics emphasize different error types. For instance, orthography (ORTH) edits, such as case changes and whitespace adjustments, tend to be downplayed. Note that such a bias in the metrics is not necessarily a bad thing. By introducing this bias, it is possible that the reference-free evaluation has improved its alignment with human evaluations.

5.2 Precision per Error Type

While the analyses so far have discussed general attribution results, here we investigate attribution results specific to GEC models. Typically, metrics with low explainability provide only a single numerical score at the corpus level. We decompose this score is into performance across different error types via our attribution. Specifically, we treat edits with positive attribution scores as True Positives, and those with negative attribution scores as False Positives, enabling the calculation of precision for each error type. To handle attribution scores across multiple sentences, we use normalized Shapley values:

$$\text{Precision} = \frac{\phi_{+}^{\text{norm}}(M)}{\phi_{+}^{\text{norm}}(M) + |\phi_{-}^{\text{norm}}(M)|}, \quad (6)$$

where $\phi_{+}^{\text{norm}}(M)$ and $\phi_{-}^{\text{norm}}(M)$ represent the sum of positive and negative normalized attribution scores at the corpus-level, respectively.

Figure 6 shows the precision for each error type using the JFLEG validation set and SOME as the evaluation metric. The parentheses in the y-axis labels indicate the corpus-level scores, with each row of the heatmap explaining these scores in terms of error types. By analyzing precision by error type, we can see that for GPT-4o-mini, edits related to adverbs (ADV) and orthography (ORTH) contribute relatively highly to the score. This indicates that errors involving these error types are play into GPT-4o mini's strengths. On the other hand, despite achieving the highest corpus-level score among the five systems, GPT-4o mini's precisions are not particularly high. Notably, T5 appears to perform better in terms of precision, as indicated by more dark-colored cells. This discrepancy may stem from an overcorrection issue, leading to a low-precision, high-recall trend in performance (Fang et al., 2023; Omelianchuk et al., 2024). While this trend is intuitive in the reference-based evaluation because the valid edits in it are limited to the references, we also observed a similar trend even for reference-free evaluation metrics.

6 Conclusion

This paper proposes a method to improve the explainability of existing low-explainable GEC met-

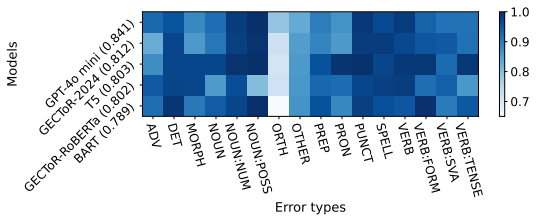


Figure 6: The heatmap indicating the precision for each GEC systems. We used JFLEG validation set as a dataset and SOME as a metric.

rics by attributing sentence-level scores to individual edits. Specifically, we employed Shapley values to perform attribution while accounting for various contexts in which edits are applied. The quantitative analysis indicates that the sign (positive or negative) of the attribution score has approximately 70% agreement rate with the correctness or incorrectness of edit-level human evaluations. We demonstrated through case studies that metric judgments can be displayed at the edit level, and analyzed them broadly as biases based on error type.

Limitations

Treating False Negative Corrections. The proposed method is limited to analyzing corrections made by the GEC system, i.e. True Positives (TP) and False Positives (FP), and does not address False Negatives (FN). Possibly, FN can be inferred by performing error detection, but we cannot apply our attribution unless it is treated as an "edit" containing the corrected string, thus it is not easy to treat FN. One solution can be considered is that the use of reference sentences, but it loses the advantage that a reference-free metric does not require reference sentences. In the proposed method, we assume that the effect of FN is canceled out by $\Delta M(H|S) = M(H|S) - M(S|S)$ because FN is included in both S and H. Thus FN does not affect the computation of attribution scores for TP and FP. A more detailed investigation into this issue is left for future work.

Treating dependent edits Edits might exhibit dependencies. For example, the correction [*model 's prediction -> prediction of the model*] can be split into two dependent edits: [*model 's -> ϕ*] and [*ϕ -> of the model*]. Although multiple corrections with such dependencies should be applied or not applied together in the process of computing the Shapley values, this study treats all edits independently. One difficult point is that there is no dataset to which the dependencies of edits are annotated, and no tools to identify edit dependencies in the current GEC field. Therefore, it is difficult to handle dependencies with the current technology. Note that CLEME (Ye et al., 2023) addressed the correction independence assumption, and they have actually succeeded in their evaluation metric that treats corrections independently. Their results suggest the validity of treating corrections independently in our study.

Rectifying Metric Biases The case study results (Section 5.1) revealed that metrics exhibit biases towards specific error types. While one could attempt to mitigate such biases, we believe that sentence-level metrics benefit from implicitly weighting edits, making these biases beneficial for evaluation. However, biases related to social factors such as gender or nationality, should be resolved. A deeper investigation into metric biases is beyond the scope of this work, but remains an important area for future research. Our work provides a strong foundation for exploring these biases.

References

Riadh Belkebir and Nizar Habash. 2021. Automatic error type annotation for Arabic. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 596–606, Online. Association for Computational Linguistics.

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.

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.

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.

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. *Preprint*, arXiv: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.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *Preprint*, arXiv:2304.01746.

Mariano Felice, Christopher Bryant, and Ted Briscoe. 2016. Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835, Osaka, Japan. The COLING 2016 Organizing Committee.

Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437.

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.

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.

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

Md Asadul Islam and Enrico Magnani. 2021. Is this the end of the gold standard? a straightforward referenceless 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.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024a. Large language models are state-of-the-art evaluator for grammatical error correction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 68–77, Mexico City, Mexico. Association for Computational Linguistics.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024b. Revisiting meta-evaluation for grammatical error correction. *Transactions of the Association for Computational Linguistics*, 12:837–855.

Katerina Korre, Marita Chatzipanagiotou, and John Pavlopoulos. 2021. ELERRANT: Automatic grammatical error type classification for Greek. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 708–717, Held Online. INCOMA Ltd.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

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.

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.

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.

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.

Kostiantyn Omelianchuk, Andrii Liubonko, Oleksandr Skurzhanskyi, Artem Chernodub, Oleksandr Korniienko, and Igor Samokhin. 2024. Pillars of grammatical error correction: Comprehensive inspection

of contemporary approaches in the era of large language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 17–33, Mexico City, Mexico. Association for Computational Linguistics.

V Petsiuk. 2018. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

Lloyd S Shapley et al. 1953. A value for n-person games.

Alexey Sorokin. 2022. Improved grammatical error correction by ranking elementary edits. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11416–11429, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Erik Strumbelj and Igor Kononenko. 2010. An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res.*, 11:1–18.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

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.

Harun Uz and Gülşen Eryiğit. 2023. Towards automatic grammatical error type classification for Turkish. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 134–142, Dubrovnik, Croatia. Association for Computational Linguistics.

Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. 2024. Gradient based feature attribution in explainable ai: A technical review. *Preprint*, arXiv:2403.10415.

Helen Yannakoudakis, Øistein E Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for esl learners. *Applied Measurement in Education*, 31(3):251–267.

Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng, and Ying Shen. 2023. CLEME: Debiasing multi-reference evaluation for grammatical error correction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6174–6189, Singapore. Association for Computational Linguistics.

Jingheng Ye, Zishan Xu, Yinghui Li, Xuxin Cheng, Linlin Song, Qingyu Zhou, Hai-Tao Zheng, Ying Shen, and Xin Su. 2024. Cleme2.0: Towards more interpretable evaluation by disentangling edits for grammatical error correction. *Preprint*, arXiv:2407.00934.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Computation Costs

Figure 7 shows the relationship between the number of edits in a sentence and its computation cost to compute attribution scores. This includes the results of both exact Shapley values and Shapley sampling values, for the metrics introduced in Section 4.1.3. In exact Shapley values, the computation takes more than 30 seconds when the number of edits exceeds 11 edits. In contrast, Shapley sampling values reduces these times to less than five seconds. For each metric, the lines for the exact Shapley values and the Shapley sampling values intersect at $N = 9$. This reason is that the number of samples to be evaluated will be almost the same; $NT = 9 * 64 = 576$ for sampling values, and $2^N = 2^9 = 512$ for the exact values.

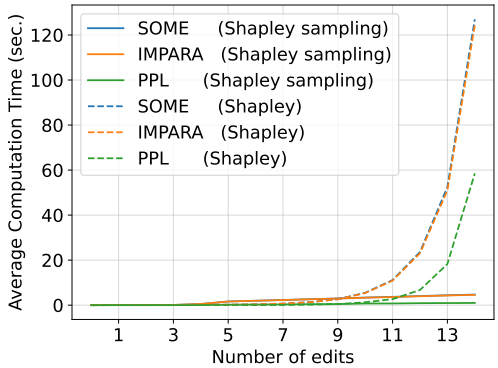


Figure 7: The relationship between the number of edits and computation time per sentence. The solid lines are average time and ranges are standard deviation.