Rethinking Evaluation Metrics for Grammatical Error Correction: Why Use a Different Evaluation Process than Human?

Takumi Goto, Yusuke Sakai, Taro Watanabe
Nara Institute of Science and Technology
{goto.takumi.gv7, sakai.yusuke.sr9, taro}@is.naist.jp

Abstract

One of the goals of automatic evaluation metrics in grammatical error correction (GEC) is to rank GEC systems such that it matches human preferences. However, current automatic evaluations are based on procedures that diverge from human evaluation. Specifically, human evaluation derives rankings by aggregating sentence-level relative evaluation results, e.g., pairwise comparisons, using a rating algorithm, whereas automatic evaluation averages sentence-level absolute scores to obtain corpus-level scores, which are then sorted to determine rankings. In this study, we propose an aggregation method for existing automatic evaluation metrics which aligns with human evaluation methods to bridge this gap. We conducted experiments using various metrics, including edit-based metrics, n-gram based metrics, and sentence-level metrics, and show that resolving the gap improves results for the most of metrics on the SEEDA benchmark. We also found that even BERT-based metrics sometimes outperform the metrics of GPT-4. The proposed ranking method is intergrated GEC-METRICS[1].

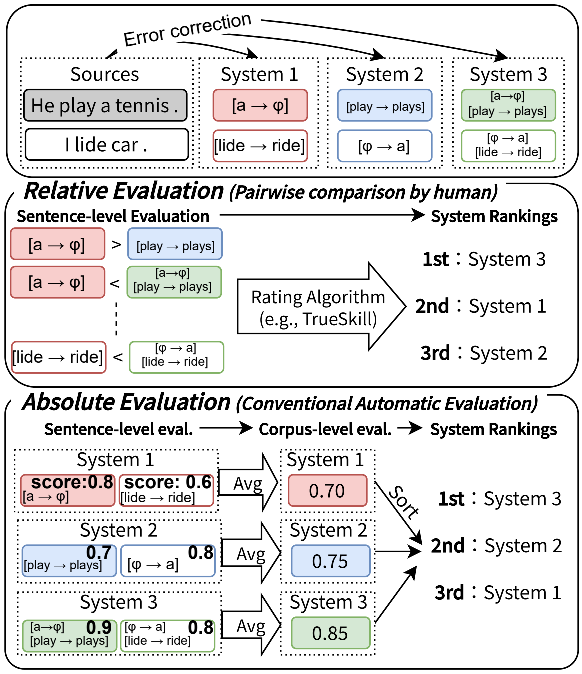


Figure 1: An overview of current human and automatic evaluation when ranking three GEC systems based on a dataset containing two sentences. Each system output represents edits for simplicity.

1 Introduction

Grammatical error correction (GEC) task aims to automatically correct grammatical errors and surface errors such as spelling and orthographic errors in text. Various GEC systems have been proposed based on sequence-to-sequence models (Katsumata and Komachi, 2020; Rothe et al., 2021), sequence tagging (Awasthi et al., 2019; Omelianchuk et al., 2020), and language models (Kaneko and Okazaki, 2023; Loem et al., 2023), and it is crucial to rank those systems based on automatic evaluation metrics to select the best system matching user's demands. Automatic evaluation is expected to rank GEC systems aligning with human preference, as evidenced by meta-evaluations of automatic metrics that assess their agreement with human evaluation (Grundkiewicz et al., 2015; Kobayashi et al., 2024b). For example, one can compute Spearman's rank correlation coefficient between the rankings produced by automatic and human evaluation, considering a metric with a higher correlation as a better metric.

However, despite the clear goal of reproducing human evaluation, current automatic evaluation is based on procedures that diverge from human evaluation. Figure 1 illustrates the evaluation procedure for ranking three GEC systems using a dataset comprising two sentences. In human evaluation, corrected sentences generated for the same input sentence are compared relatively across sys-

[1]A library for GEC evaluation proposed by Goto et al. (2025), https://github.com/gotutiyan/gec-metrics.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1165–1172
July 27 - August 1, 2025

tem outputs, i.e., pairwise comparison, and the results are aggregated as rankings using rating algorithms such as TrueSkill (Herbrich et al., 2006). In contrast, automatic evaluation estimates sentence-wise scores, then averages them at the corpus level and determines rankings by sorting these averaged scores. As such, current automatic evaluation follows a procedure that deviates from human evaluation, contradicting the goal of reproducing human judgment. Intuitively, it would be desirable for automatic evaluation to follow the same procedure as human evaluation.

In this study, we hypothesize that resolving this gap will more closely align automatic evaluation with human evaluation. Based on this hypothesis, we propose computing rankings in automatic evaluation using the same procedure as human evaluation, e.g., using TrueSkill after deriving pairwise estimates based on sentence-wise scores when human evaluation employs TrueSkill. In our experiments, we conducted a meta-evaluation on various existing automatic evaluation metrics using the SEEDA dataset (Kobayashi et al., 2024b), that is a representative meta-evaluation benchmark. The results show that bridging the identified gap improves ranking capability for many metrics and that BERT-based (Devlin et al., 2019) automatic evaluation metrics can even outperform large language models (LLMs), GPT-4 (OpenAI et al., 2024), in evaluation. Furthermore, we discuss the use and development of automatic evaluation metrics in the future, emphasizing that sentence-level relative evaluation is particularly important for developing new evaluation metrics.

2 Gap Between Human and Automatic Evaluation

2.1 Background

Human evaluation has been conducted by Grundkiewicz et al. (2015), who manually evaluated systems submitted to the CoNLL-2014 shared task (Ng et al., 2014), and by Kobayashi et al. (2024b), who included state-of-the-art GEC systems such as LLMs in their dataset. In both studies, system rankings were derived by applying a rating algorithm to sentence-level pairwise comparisons. Commonly used rating algorithms include Expected Wins (Bojar et al., 2013) and TrueSkill (Herbrich et al., 2006; Sakaguchi et al., 2014). Grundkiewicz et al. (2015) adopted Expected Wins as their final ranking method, whereas Kobayashi et al. (2024b) used TrueSkill to determine the final ranking. Kobayashi et al. (2024b) also pointed out the importance of aligning the granularity of evaluation between automatic evaluation and human evaluation, but did not mention the procedure for converting sentence-level evaluation into system rankings.

Automatic evaluation is conducted using various evaluation metrics, including reference-based and reference-free approaches, as well as sentence-level and edit-based metrics. Most of these metrics follow a procedure in which each sentence is assigned an absolute score, which is then aggregated into a corpus-level evaluation score. For example, sentence-level metrics such as SOME (Yoshimura et al., 2020) and IMPARA (Maeda et al., 2022) aggregate scores by averaging, while edit-based metrics such as ERRANT (Felice et al., 2016; Bryant et al., 2017) and GoToScorer (Gotou et al., 2020), as well as n-gram-based metrics such as GLEU (Napoles et al., 2015, 2016) and GREEN (Koyama et al., 2024), aggregate scores by accumulating the number of edits or n-grams. The corpus-level scores obtained through these methods can be converted into system rankings by sorting.

2.2 How to Resolve the Gap?

The gap can be simply addressed by using automatic evaluation metrics in the same manner as human evaluation. Given that the SEEDA dataset uses TrueSkill as the aggregation method, we will close the gap by using TrueSkill for automated evaluation as well. First, since existing automatic evaluation metrics compute sentence-wise scores, we convert these scores into pairwise comparison results. For example, in the case illustrated in Figure 1, the evaluation scores of 0.8, 0.7, and 0.9 for corrected sentences corresponding to the first sentence ("He play a tennis") can be compared to produce pairwise comparison results similar to those in human evaluation. Next, we compute system rankings by applying TrueSkill to the transformed pairwise comparison results. In this study, we consider all combinations of pairwise comparisons for system set. That is, given N systems, a total of $N(N-1)$ comparisons are performed per sentence, and system rankings are computed based on these results including ties.

A similar method was employed by Kobayashi et al. (2024a), but they did not mention the gap. Also, their experiments used the TrueSkill aggregation for their proposed LLM-based metrics, but used conventional aggregation methods, e.g., averaging, for other metrics. We discuss and organize the gap between human and automatic evaluation in detail, and then solve the gap by applying TrueSkill to all metrics for fair comparison.

Note that our method is explained using TrueSkill, which is used as the human evaluation method for SEEDA. If another meta-evaluation dataset uses a different aggregation method, such as Expected Wins, we should use that instead. We emphasize that our claim is the importance of aligning the aggregation methods of human and automatic evaluations, as even this simple practice has been largely overlooked so far.

3 Experiments

3.1 Automatic Evaluation Metrics

We provide more detailed experimental settings for each metric in Appendix A.

Edit-based metrics We use ERRANT (Felice et al., 2016; Bryant et al., 2017) and PT-ERRANT (Gong et al., 2022). Both are reference-based evaluation metrics that assess at the edit level. When multiple references are available, the reference that yields the highest $F_{0.5}$ score is selected for each sentence.

n-gram based metrics We use GLEU+ (Napoles et al., 2015, 2016) and GREEN (Koyama et al., 2024). The n-gram overlap is checked among the input sentence, hypothesis sentence, and reference sentence. When multiple references are available, GLEU+ uses the average score across all references, and GREEN uses the reference that yields the highest score is selected for each sentence.

Sentence-level metrics SOME (Yoshimura et al., 2020), IMPARA (Maeda et al., 2022), and Scribendi Score (Islam and Magnani, 2021) are used. All of them are based on small neural models such as $\text{BERT}_{\text{base}}$ (Devlin et al., 2019) and designed as a reference-free metric that considers the correction quality estimation score as well as the meaning preservation score between the input and corrected sentences.

3.2 Meta-Evaluation Method

We use the SEEDA dataset (Kobayashi et al., 2024b) for meta-evaluation. Meta-evaluation results are reported based on human evaluation results using TrueSkill for both the sentence-level human-evaluation, SEEDA-S, and the edit-level human-evaluation, SEEDA-E. Additionally, we also report results for both the `Base` configuration, which excludes the fluent reference and GPT-3.5 outputs that allow for larger rewrites, and the `+Fluency` configuration, which includes them.

Furthermore, we evaluate the robustness of the calculated rankings using window analysis (Kobayashi et al., 2024a). The window analysis computes correlation coefficients only for consecutive N systems, after sorting systems based on human evaluation results. This allows us to analyze whether automatic evaluation can correctly assess a set of systems that appear to have similar performance from the human evaluation. In this study, we perform it with $N = 8$ for 14 systems corresponding to the `+Fluency` configuration, and report both Pearson and Spearman correlation coefficients. That is, correlation coefficients are computed for the rankings 1 to 8, 2 to 9, . . ., and 7 to 14 from human evaluation.

3.3 Experimental Results

Table 1 shows the results of the meta-evaluation. The upper group presents evaluation results based on the conventional method of averaging or summing, and the bottom group presents results evaluated using TrueSkill, which follows the same evaluation method as human evaluation. The bottom group includes the evaluation results based on GPT-4 reported by Kobayashi et al. (2024a), which correspond to the state-of-the-art metrics.

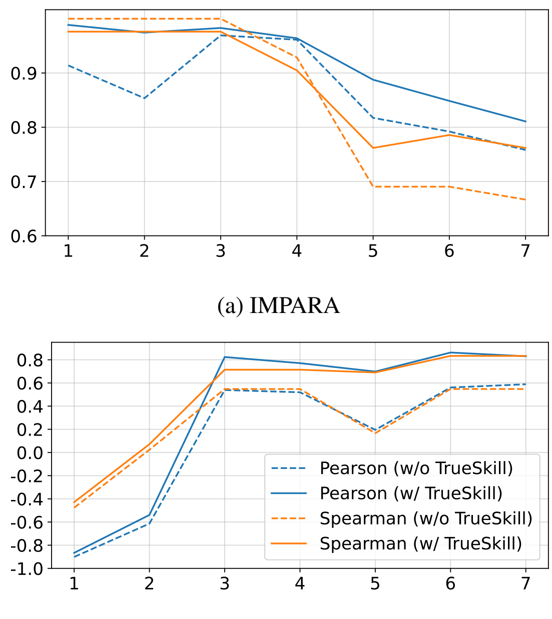
The overall trend indicates that using TrueSkill-based evaluation improves the correlation coefficients for most of metrics. In particular, the results of IMPARA in the SEEDA-S and `+Fluency` setting outperformed those of GPT-4 results. Additionally, ERRANT showed an improvement of more than 0.2 points in many configurations. These results show that using automatic evaluation metrics with the same evaluation procedure as human evaluation makes the ranking closer to human evaluation. In other words, the existing automatic evaluation metrics were underestimated in the prior reports due to the gap in the meta-evaluation procedure.

Unlike edit-level or sentence-level metrics, no effect was observed in n-gram-level metrics such as GLEU+ and GREEN. This stems from the poor sentence-level evaluation performance of n-gram based metrics. The BLEU paper (Papineni et al., 2002), which is a n-gram-level metric for machine translation and the basis of the GLEU+, notes that the brevity penalty can excessively penalize scores for short sentences, and uses corpus-level lengths

Metrics	SEEDA-S Base r (Pearson)	SEEDA-S Base ρ (Spearman)	SEEDA-S `+Fluency` r	SEEDA-S `+Fluency` ρ	SEEDA-E Base r	SEEDA-E Base ρ	SEEDA-E `+Fluency` r	SEEDA-E `+Fluency` ρ
w/o TrueSkill								
ERRANT	0.545	0.343	-0.591	-0.156	0.689	0.643	-0.507	0.033
PTERRANT	0.700	0.629	-0.546	0.077	0.788	0.874	-0.470	0.231
GLEU+	0.886	0.902	0.155	0.543	0.912	0.944	0.232	0.569
GREEN	0.925	0.881	0.185	0.569	0.932	0.965	0.252	0.618
SOME	0.892	0.867	0.931	0.916	0.901	0.951	0.943	0.969
IMPARA	0.916	0.902	0.887	0.938	0.902	0.965	0.900	0.978
Scribendi	0.620	0.636	0.604	0.714	0.825	0.839	0.715	0.842
w/ TrueSkill								
ERRANT	<u>0.763</u>	<u>0.706</u>	<u>-0.463</u>	<u>0.095</u>	<u>0.881</u>	<u>0.895</u>	<u>-0.374</u>	<u>0.231</u>
PTERRANT	<u>0.870</u>	<u>0.797</u>	<u>-0.366</u>	<u>0.182</u>	<u>0.924</u>	<u>0.951</u>	<u>-0.288</u>	<u>0.279</u>
GLEU+	0.863	0.846	0.017	0.393	0.909	<u>0.965</u>	0.102	0.486
GREEN	0.855	0.846	-0.214	0.327	0.912	0.965	-0.135	0.420
SOME	<u>0.932</u>	<u>0.881</u>	<u>0.971</u>	<u>0.925</u>	0.893	0.944	<u>0.965</u>	0.965
IMPARA	<u>0.939</u>	<u>**0.923**</u>	<u>**0.975**</u>	<u>**0.952**</u>	0.901	0.944	<u>0.969</u>	0.965
Scribendi	<u>0.674</u>	<u>0.762</u>	<u>0.745</u>	<u>0.859</u>	<u>0.837</u>	<u>0.888</u>	<u>0.826</u>	<u>0.912</u>
GPT-4-E (fluency)	0.844	0.860	0.793	0.908	0.905	**0.986**	0.848	0.987
GPT-4-S (fluency)	0.913	0.874	0.952	0.916	**0.974**	0.979	**0.981**	**0.982**
GPT-4-S (meaning)	**0.958**	0.881	0.952	0.925	0.911	0.960	0.976	0.974

Table 1: Correlation with human evaluation using the SEEDA dataset. *w/o TrueSkill* refers to the conventional evaluation procedure, while *w/ TrueSkill* represents the proposed evaluation procedure. Improvements over the conventional procedure are <u>underlined</u>, and the highest value in each column is highlighted in **bold**. The GPT-4 results refer to those reported in Kobayashi et al. (2024b).

to address this issue. Since GLEU+ also employs a brevity penalty, it cannot accurately calculate sentence-level scores depending on sentence length. This is a serious issue for TrueSkill-based aggregation because the quality of sentence-level scores directly affects the quality of system rankings. Furthermore, while GREEN does not use a brevity penalty, the score can become unstable as the "n" for n-gram increases, especially for short sentences. This has a negative impact on the geometric mean among n-gram scores, which is the final score. An ideal metric should provide evaluation results that better align with human judgments when ranking systems in the same way humans do. Given this premise, our results suggest a human alignment issue of n-gram-level metrics.



(a) IMPARA

(b) ERRANT

Figure 2: The results of the window analysis for $N = 8$ are shown. The x-axis represents the starting rank of human evaluation. For example, $x = 2$ shows the results for the systems ranked 2nd to 10th in human evaluation.

Figure 2 shows the results of the window analysis for IMPARA and ERRANT measured on SEEDA-S and SEEDA-E, respectively. From Figure 2a, it can be seen that IMPARA particularly aligns with human evaluation in the lower ranks. The Pearson correlation coefficient also showed an improvement in the evaluation results for the top systems as well. Since the top systems include GEC systems that are largely rewritten, such as GPT-3.5, this characteristic is useful, considering that LLM-based correction methods will become popular in the future. Figure 2b shows that ERRANT consistently showed improved correlation coefficients with the proposed method, but still struggled with evaluating the top systems. For edit-based evaluation metrics, it is still considered difficult to assess such GEC systems even with the evaluation method aligned with human evaluation[2].

4 Conclusion

In this study, we focused on the fact that human evaluation aggregates sentence-level scores into system rankings based on TrueSkill, while automatic evaluation uses a different evaluation, and we proposed to use TrueSkill in automatic evaluation as well. Experimental results with various existing metrics showed improvements in correlations with human evaluation for many of the metrics, indicating that agreement on the aggregation method is important. Our core statement in this paper is not about using TrueSkill, but rather the importance of using the same aggregation method as human evaluation. For instance, if future meta-evaluation datasets switch the aggregation method for human evaluation to averaging sentence-level scores, then automatic evaluation should likewise adopt the same approach.

Given the discussion so far, at least in the current situation of GEC evaluation, we recommend transitioning the aggregation method from averaging or summing to using a rating algorithm, such as TrueSkill. We also recommend that evaluation metrics should be developed that allow for accurate sentence-wise comparisons. This is evidenced by the fact that IMAPARA achieves a higher correlation coefficient than SOME in Table 1. In fact, IMAPARA is trained to assess the pairwise comparison results, whereas SOME is trained to evaluate sentences absolutely.

Limitations

Use for Purposes Other Than System Ranking The proposed method is designed for system ranking and cannot be used for other types of evaluation, such as analyzing the strengths and weaknesses of a specific system. For instance, when analyzing whether a model excels in precision or recall, it is more useful to accumulate the number of edits at the corpus level, as done in existing evaluation methods.

Reproducing the Outputs of Compared GEC Systems Since the proposed ranking method requires inputting all GEC outputs being compared, it is necessary to reproduce their models. This point is different from existing absolute evaluation methods, where previously reported scores can be cited. While this may seem burdensome for researchers, it can also be seen as an important step toward promoting the publication of reproducible research results.

[2] Using a larger number of references may solve this issue.

Ethical Considerations

When the metric contains social biases, the proposed method cannot eliminate that bias and may reflect that bias in the rankings. However, we argue that this problem should be resolved as a metric problem.

Acknowledgments

We sincerely thank the anonymous reviewers for their insightful comments and suggestions. This work has been supported by JST SPRING. Grant Number JPMJSP2140.

References

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.

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.

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.

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.

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.

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 Goto, Yusuke Sakai, and Taro Watanabe. 2025. gec-metrics: A unified library for grammatical error correction evaluation. *Preprint*, arXiv:2505.19388.

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.

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.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. Trueskill™: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

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.

Masahiro Kaneko and Naoaki Okazaki. 2023. Reducing sequence length by predicting edit spans with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10017–10029, Singapore. Association for Computational Linguistics.

Satoru Katsumata and Mamoru Komachi. 2020. Stronger baselines for grammatical error correction using a pretrained encoder-decoder model. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 827–832, Suzhou, China. 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.

Shota Koyama, Ryo Nagata, Hiroya Takamura, and Naoaki Okazaki. 2024. n-gram F-score for evaluating grammatical error correction. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 303–313, Tokyo, Japan. Association for Computational Linguistics.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of GPT-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 205–219, Toronto, Canada. Association for Computational Linguistics.

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, 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.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2016. Gleu without tuning. *Preprint*, arXiv:1605.02592.

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.

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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 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.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

A Detailed Experimental Settings for Evaluation Metrics

We used GEC-METRICS (Goto et al., 2025) for the implementation. The detailed experimental settings are as follows:

ERRANT Evaluations were conducted using the Python module `errant==3.0.0` with the span-based correction setting. Although the edits in the CoNLL-2014 references were manually annotated, they were re-extracted using ERRANT. This is important for an evaluation based on consistent edits.

PT-ERRANT We employed the $F1$ score of BERTScore for the weighting. The baseline rescaling was adopted, and IDF adjustment was not performed. We used `bert-base-uncased` for the BERT model. These settings are consistent with those of the official implementation [3]. Similar to ERRANT, we re-extracted reference edits via `errant` module.

GLEU+ We used **GLEU+** with n-grams up to 4-grams and 500 iterations during reference sampling.

GREEN Similarly, we used **GREEN** with n-grams up to 4-grams, utilizing the $F_{2.0}$ score, following those employed by Koyama et al. (2024).

SOME We used the official models for grammaticality, fluency, and meaning preservation, with respective weights of 0.55, 0.43, and 0.02. These weights correspond to those tuned by Yoshimura et al. (2020) for sentence-level evaluation performance.

IMAPARA As a pre-trained quality estimation model was not publicly available, we newly constructed it through reimplementation and experimentation. Following Maeda et al. (2022), the CoNLL-2013 dataset (Ng et al., 2013) was used as a seed corpus and was split into training, development, and evaluation sets with an 8:1:1 ratio. We fine-tuned `bert-base-cased`, and followed `BertForSequenceClassification` from the Transformers library [4] (Wolf et al., 2020) for the classifier architecture. This corresponds to transforming the `CLS` representation of the BERT model into a real value via a single projection layer. During inference, `bert-base-cased` was

[3] https://github.com/pygongnlp/PT-M2
[4] https://github.com/huggingface/transformers

used for the similarity estimation model, with a threshold set to 0.9.

Scribendi Score GPT-2 (Radford et al., 2019)[5] was utilized as the language model, and the threshold for the maximum of the Levenshtein-distance ratio and token sort ratio was set to 0.8.

[5] https://huggingface.co/openai-community/gpt2